TRABAJO DE FIN DE GRADO
Grado en Ingeniería Informática

Herramienta visual para monitorización tweets durante eventos excepcionales

---

Autor:
Belén Carrión Recio

Tutor:
Teresa Onorati

Resumen

Vivimos en una sociedad en las que las redes sociales juegan un papel muy importante en nuestras vidas. Estas nos permiten mantenernos en contacto con nuestros seres queridos, nos dan un lugar en el que expresar nuestros sentimientos, o compartir lo que nos encontramos haciendo. La popularidad de estas redes está en auge, con una enorme cantidad de contenido nuevo publicándose cada segundo.

El objetivo de este Trabajo de Fin de Grado es desarrollar una herramienta que permita dar un significado a esta información, analizándola y visualizándola, de modo que pueda ser útil para los servicios de emergencia. Estos, de momento, no emplean esta información, y usarla podría suponer un avance en su labor.

Tras un estudio inicial sobre cómo las redes sociales han sido empleadas por otros investigadores en el ámbito de los servicios de emergencia, así como un estudio de las herramientas de visualización de tweets existentes, se propone la siguiente herramienta visual para monitorización de tweets durante eventos excepcionales.

Esta herramienta recupera información de la red social Twitter y la almacena en una base de datos. Los tweets - estados compartidos en esta red - son analizados sintáctica y semánticamente y categorizados en función de su relevancia.

Los tweets se muestran sobre un mapa en las coordenadas en las que fueron escritos y agrupados en clusters - grupos con coordenadas próximas. Cada cluster es representado por el polígono de menor área que recubra todos los marcadores del cluster. Se calcula el término más frecuente entre los tweets que pertenecen al cluster. La categoría de este es la que determina el color del polígono. Este término más frecuente también se muestra sobre el mapa.

En último lugar, se aplica la herramienta a un caso de uso real: los atentados de París del 13 de noviembre de 2015, y se presentan una serie de conclusiones.

Abstact

We live in a society where social networks play a very important role in our lives. They allow us to keep in touch with the people we care about, they give us a place where we can express our feelings, or share what we are doing. The popularity of these networks keeps rising, with a high amount of new content being published every second.

This project aims to develop a tool that will make use of all this information, analyze it and visualize it in a way that it can become useful for the emergency systems. These, at the moment, are not using this information, and using it could lead to an improvement in their performance.

After a prior study on how social networks have been used by other researchers in the field of emergency systems, as well as a study on the existing tweet visualization tools, a visual tool for monitoring tweets during exceptional events is proposed.

This tool retrieves data from the social network Twitter and stores it in a database. The tweets - statuses shared on this network - are later syntactically and semantically analyzed and categorized based on their relevance.

Tweets are shown on a map on the coordinates where they were shared, and clustered into groups with nearby coordinates. Each cluster is represented by the polygon with the minimum area that covers all the markers in the polygon. The most used term in each cluster is calculated. The category this term belongs in determines the color of the polygon. This most frequent term is also shown on the map.

Lastly, the tool is tested using a real use case: the Paris attacks of November 13th, 2015, and some conclusions are presented.

Índice

Índice de figuras

Índice de tablas

1. Introducción

Este Trabajo de Fin de Grado se centra en la implementación de una herramienta de visualización de tweets para la monitorización de la actividad en las redes sociales durante eventos excepcionales. En esta sección, se describen la motivación de este proyecto, los objetivos que se desean conseguir, el problema del que se parte y la solución propuesta para solucionarlo.

1.1. Entorno socio-económico

En los últimos años el uso de las tecnologías ha ido en aumento. Se trata de una sociedad en la que la edad media a la que una persona adquiere su primer teléfono móvil es de 13 años [1], y el 80 % de la población de nuestro país cuenta con un smartphone [2]. El uso de teléfonos móviles esta ciertamente arraigado.

A diferencia de los teléfonos fijos, estos nos permiten mantenernos en contacto con otras personas en cualquier momento y desde cualquier lugar. Lo cual tiene un gran impacto tanto en la vida cotidiana como en el ámbito laboral.

Además, los smartphones cuentan con una gran cantidad y variedad de aplicaciones: desde juegos para entretenerse, calculadora, notas, o reproductores de música. Otra gran ventaja que ofrece es el acceso a internet, que simplifica más aún el poder mantenerse en contacto, por ejemplo, a través de las redes sociales.

Se ha visto también un aumento en el uso de estas. Se trata de plataformas que permiten la comunicación y el contacto entre un gran número de usuarios. En éstas se pueden encontrar mensajes de tipos muy distintos. Pueden emplearse para compartir lo que nos encontramos haciendo, o compartir pensamientos y opiniones. También es habitual que se publiquen noticias a través de ellas.

La aparición de las redes sociales ha facilitado la comunicación de distintas formas. Por un lado, ofrecen un canal de comunicación, fácil de usar y disponible en todo momento. Por otro, la mayoría de ellas facilitan a sus usuarios encontrar a personas que podrían conocer: antiguos compañeros del colegio, colegas de profesión, familiares, o cuentas asociadas a números de teléfonos guardados en la agenda del usuario.

Estas redes también permiten compartir contenido y contactar con personas desconocidas, haciendo las publicaciones que en ellas se escriben de acceso público, de modo que puedan ser leídos por cualquier persona que llegue a ellas, incluyendo los que no estén registrados en la red social en cuestión. Por tanto, permiten llegar a un amplio número de lectores, tanto conocidos como desconocidos.

El impacto de estas plataformas en la sociedad es realmente interesante. Por ejemplo, la plataforma Facebook cuenta con 1,65 billones de usuarios activos al mes [3], de los cuales 1,51 billones están activos a través de los teléfonos móviles.

Esto es más que la población del país más poblado del planeta, China, que cuenta con 1,38 billones de habitantes [4]. Por tanto, si Facebook fuese un país, sería el mayor del mundo. Teniendo en cuenta que la población global es de 7.4 billones [5], 2 de cada 7 habitantes residirían en el país de Facebook.

Si comparamos la población de los países más poblados con los usuarios activos de las principales redes sociales, obtenemos los siguientes datos (Figura 1):

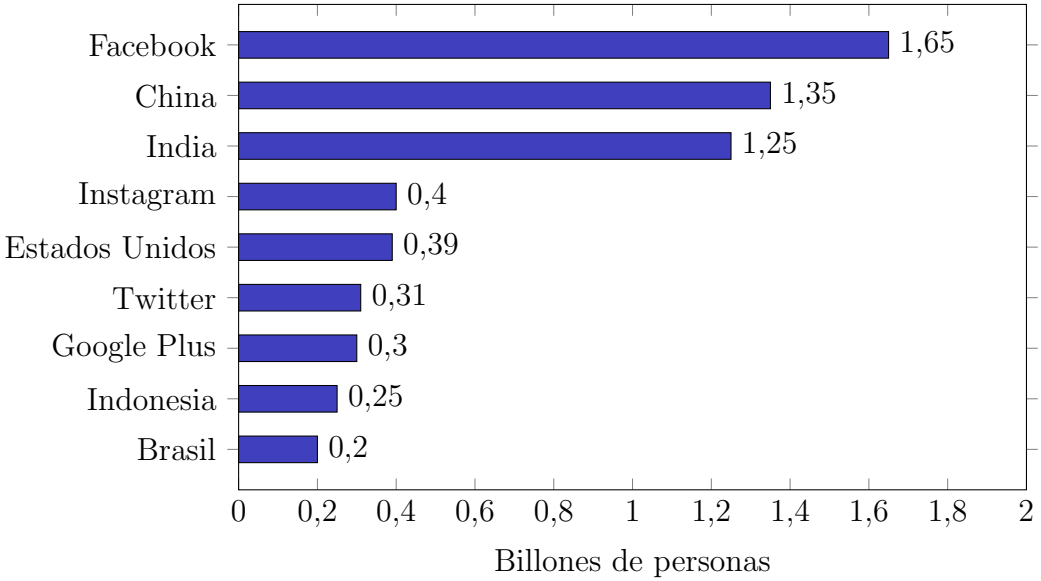


Figura 1: Población de los países más poblados comparada con los usuarios activos de las principales redes sociales

Con estos datos se demuestra la alta tasa de penetración de las redes sociales en la sociedad actual.

Por otro lado, durante el corriente año (2015-2016), Europa ha sido víctima de dos grandes ataques yihadistas en las capitales de Francia y Bélgica. Estos tipos de ataques no son algo nuevo, sino que vienen ocurriendo frecuentemente a lo largo de los últimos años. Ya en el año 2001 fue la ciudad de Nueva York víctima de uno de ellos, sin olvidar los atentados de Madrid del 11 de marzo de 2004. Una amenaza que preocupa a gran parte de la población, y que como tal genera gran respuesta en las redes sociales, cuya actividad aumenta cuando se producen eventos de este alcance.

Es habitual que los ciudadanos acudan a las redes sociales con diversos fines: en busca de información, para intentar contactar con sus seres queridos, expresar su preocupación, o para expresar muestras de apoyo a las víctimas.

No obstante, este tipo de eventos no es el único que agita las redes sociales. Otros eventos con efectos similares son las catástrofes naturales - como el huracán Sandy, que sacudió en 2014 a los Estados Unidos de América - o celebraciones especiales - como por ejemplo, algunas competiciones deportivas.

1.2. Motivación

Como se ha mencionado anteriormente en este documento, existe una gran cantidad de usuarios que acuden a las redes sociales para publicar contenido de todo tipo. Esta actividad, además, crece durante circunstancias excepcionales.

Sería interesante para un operario de un servicio de emergencia comprender qué está ocurriendo entre la población en el momento en que se desatan estos eventos. Cuando un accidente o una emergencia ocurre, es de vital importancia que estos operadores puedan responder lo más rápidamente posible. Para ello, necesitan tener toda la información necesaria y actualizada. Con tantos usuarios compartiendo tanta información, las redes sociales se convierten en un lugar del que extraer información de primera mano, directamente de la población.

1.3. Problema

Pese a que la información se encuentra ya en las redes sociales, es necesario procesarla y visualizarla de alguna manera. En un solo segundo se comparten más de 7.000 tweets [6]. Sería inconcebible para un operario leer todas las publicaciones que se están compartiendo en un instante. Puesto que se busca una respuesta rápida, detenerse a leer toda esta cantidad de publicaciones le llevaría demasiado trabajo e impediría reaccionar a tiempo.

Esto nos presenta frente al siguiente reto: ¿cómo puede esa gran cantidad de información ayudar a los servicios de emergencia?

1.4. Objetivos

El objetivo de este Trabajo de Fin de Grado es precisamente ofrecer una herramienta para que los servicios de emergencia puedan utilizar esta gran cantidad de información publicada en Twitter durante una situación excepcional y analizarla de tal forma que cobre significado y sea de utilidad para los operarios de los servicios de emergencia.

Con este objetivo en mente, se ha desarrollado una herramienta de visualización de tweets para monitorización de eventos excepcionales.

La herramienta propuesta busca alcanzar los siguientes objetivos específicos:

- Recuperar los mensajes compartidos en las redes sociales sobre un tema específico.
- Analizar semánticamente los datos recuperados, categorizándolos según su relevancia respecto al tema tratado.

- Visualizar los datos eficientemente para que los usuarios, y en particular los operadores de emergencia, puedan sacar partido de ellos.
- Aplicar la herramienta propuesta a un caso de uso real para evaluar su alcance y utilidad.

De este modo, se consigue dar un paso más en el funcionamiento tradicional de un servicio de emergencia, utilizando información ya existente y obtenida directamente de la población, para comprender mejor el alcance y el efecto de estos eventos.

1.5. Solución Final

Para resolver el problema existente - encontrar una forma de dar sentido y visualizar la gran cantidad de datos que se puede obtener de las publicaciones de los usuarios en las redes sociales - se propone la siguiente herramienta de visualización de tweets.

Twitter es una plataforma de micro-blogging creada en 2006, cuyos usuarios pueden leer y publicar mensajes con un máximo 140 caracteres. Además, tienen la posibilidad de acceder a la red social bien directamente desde su página web, o bien desde aplicaciones para smartphones y tablets, por lo que pueden contribuir con nuevo contenido fácilmente desde diversos dispositivos y en cualquier momento.

Para comprender el alcance de esta plataforma cabe destacar que en el segundo cuarto de 2015, Twitter contaba con una media de 310 millones de usuarios activos al mes [7]. Tan solo en un segundo se publica una media de 7.203 tweets [6]. Estos datos demuestran que gran parte de nuestra población se encuentra activa en esta red social, en la que se da una frecuencia de publicación de nuevo contenido alta.

Además, los usuarios de Twitter tienen la posibilidad de elegir entre crear un perfil público - cuyas publicaciones pueden ser vistas por cualquier persona, registrada o no - o privado - cuyas publicaciones sólo pueden ser vistas por los usuarios a los que ellos autoricen. Esto implica que un gran número de tweets son de público. Por todo esto, se ha decidio basar esta herramienta en datos extraídos de Twitter y no de redes sociales.

1.6. Estructura del documento

A continuación se identifican las distintas secciones de este documento, y el contenido de cada una de ellas:

1. **Estado del arte**: Se analiza y describibe el conocimiento y la situación actual en materia del empleo de redes sociales para ayudar a los servicios de emergencias.

2. **Planteamiento del problema**: Se describe cómo se ha decidido plantear la solución al problema descrito.
3. **Diseño de la solución técnica**: S describe cómo se ha diseñado la solución técnica se se ha planteado en el apartado anterior.
4. **Caso de estudio**: Se especifica el caso de uso concreto que se ha empleado para probar la herramienta, y la evaluación de esta.
5. **Planificación del trabajo**: Se detalla cómo se ha planificado el proyecto.
6. **Presupuesto**: Se identifica el presupuesto que supondría la realización de este proyecto.
7. **Conclusiones**: Se presentan las conclusiones extraídas tras la realización del proyecto.

2. Estado del arte

A lo largo de los siguientes apartados, se va a analizar y describir el conocimiento y situación actual en materia del empleo de redes sociales para ayudar a los servicios de emergencias a interpretar la información compartida.

2.1. Redes sociales y emergencias

Los usuarios de las redes sociales comparten frecuentemente lo que se encuentran haciendo, o en lo que están pensando. Durante situaciones de emergencia, es habitual que el número de publicaciones en estos sitios se disparen: supervivientes compartiendo su ubicación y pidiendo ayuda, testigos compartiendo su experiencia, mensajes de apoyo del resto de la población...

En los últimos años, varios investigadores han trabajado en encontrar formas de transformar toda esta gran cantidad de contenido en información útil, con el fin de facilitar la labor de los servicios de emergencias. Estos de momento no utilizan esta información y de otra forma no podrían procesar todo lo que está siendo compartido. Hacer uso de esta información podría suponer una mejora para su trabajo, puesto que se trata de información obtenida directamente de la población.

Uno de los primeros estudios realizados data de 2007. Se trata de Java et al. [8] y en él se describe el crecimiento de la plataforma de microblogging Twitter. Ya entonces los autores observaron que las noticias de última hora y eventos excepcionales generan un aumento de la actividad.

En 2009, Hughes y Palen [9] estudiaron el comportamiento en Twitter durante varios eventos excepcionales ocurridos en Estados Unidos, y los compararon con la actividad regular. Eligieron dos emergencias - el huracán Gustav y el huracán Ike - y dos eventos de seguridad nacional - las convenciones de los partidos Democrático y Republicano - y recopilaron la actividad que se había generado durante la duración. Pudieron observar que se la actividad aumentaba: más gente se registraba en la red social durante estos eventos, y el porcentaje de usuarios inactivos - considerando como tal a aquellos que compartiesen menos de una publicación a la semana - disminuía.

En el año 2012, el huracán Sandy golpeó la costa atlántica de los Estados Unidos. En este año, la popularidad de las redes sociales había aumentado, y tras el huracán aumenta también el número de estudios sobre el tema. Preis et al. [10] estudian el impacto en Flickr - comunidad que permite compartir fotografías -, donde aprecian un aumento en las imágenes cuyo texto esté relacionado con huracanes. Onorati y Diaz [11] desarrollan WallTweet, una herramienta de visualización de la actividad generada en Twitter durante el huracán Sandy.

2.2. Herramientas de visualización de la actividad en redes sociales

Teniendo en cuenta el alcance de las redes sociales y las posibilidades que supone tener esta gran cantidad de información accesible públicamente, surge la idea de recoger colecciones de tweets y estudiarlos con diversos fines.

Twitscoop [12] parte de una lista de temas y muestra una nube de etiquetas en la que cada tema aparece escrito con un tamaño de fuente que aumenta en proporción a la cantidad de tweets que se han encontrado con ese tema. Los autores utilizan las Tendencias - lista de términos más utilizados en Twitter en un momento concreto - proporcionadas por Twitter.

Al igual que la herramienta propuesta en este documento, TweetTracker [13] utiliza dominios de emergencias, desastres y ayuda humanitaria y muestra en un mapa un marcador en las coordenadas con las que se haya encontrado un tweet. A diferencia de la herramienta desarrollada, ésta no ofrece una visualización con animación - en la que los tweets se van agregando progresivamente en el orden en el que fueron compartidos-, no muestra los términos más frecuentes, ni realiza una agrupación de marcadores o categorización de los términos.

Partiendo de distintos algoritmos ya existentes, Gansner et al. [14] han desarrollado un nuevo algoritmo para organizar una serie de marcadores en clusters (grupos de marcadores próximos), agrupando conjuntos de tweets en función del tema que traten según su análisis sintáctico.

[15] utilizan una base de datos musical donde almacenan tweets en los que se mencionan géneros musicales. Se trata de una visualización en la que se representa una matriz de dos dimensiones encima de un mapa. En cada celda de esta matriz aparece un punto cuyo color depende del género musical más frecuente entre los tweets cuyas coordenadas pertenezcan a la celda en cuestión. Además, el tamaño de este punto varía aumentando cuantos más tweets se hayan encontrado sobre el género musical.

Tras recoger tweets sobre pandemias, [16] generan un conjunto de gráficos de barras básicos con estadísticas. [17] buscan estudiar los cambios en los patrones de la geolocalización de los tweets publicados en el campus de la universidad de Purdue a lo largo de una semana. Para ello recogen por separado los tweets publicados entre semana, y aquellos publicados durante el fin de semana y los representa en dos mapas, agrupándolos en clusters en función de la distancia entre sus coordenadas. Cada cluster es representado por un círculo, cuyo tamaño aumenta en función del número de tweets que contenga.

[18] proponen una herramienta que permite identificar noticias de última hora antes que los métodos tradicionales de reportaje. Agrupa los tweets en tres categorías - eventos naturales, eventos causados por el hombre, o eventos sin categoría - y los agrupa en clusters. Por cada cluster, muestra una etiqueta sobre el mapa, cuyo color depende de la categoría a la que pertenezcan sus tweets, y que

muestra el término más común entre los tweets del cluster.

La tabla a continuación recoge las principales características de algunas de estas herramientas.

Herramienta	**Dominio**	**Datos**	**Visualización**
Twitscoop [12]	Interés general	Tendencias	Nube de etiquetas
TweetTracker [19]	Emergencias	Desastres y ayuda humanitaria	Marcadores en mapa
Gansner et al. [14]	Interés general	Temas de interés	TwitterScope
Hauger et al. [15]	Música	Géneros musicales	Matriz sobre mapa
Chew y Eysenbach [16]	Emergencias	Pandemias	Gráfico de barras
Li y Shan [17]	Interés general	Ubicación concreta	Clusters sobre mapa
Meyer et al. [18]	Noticias	Noticias	Palabras claves en mapa
The Vox Civita [20]	Opiniones	Tweets etiquetados	Línea del tiempo
Hao et al. [21]	Sentimientos	Desconocido	Distribución geográfica
SensePlace2 [22]	Interés general	Ubicación concreta	Mapa de calor
Cao et al. [23]	Interés general	Tema concreto	Mapa de girasol
Yin et al. [24]	Emergencias	Evento concreto	Clusters y etiquetas en mapa
TweetXplorer [13]	Retweets	Usuarios y tweets	Mapa de calor populares
TopicPanorama [25]	Interés general	Tema concreto en diferentes plataformas	Gráfica circular
OpinionFlow [26]	Opiniones	Desconocido	Línea del tiempo
Matisse	Opiniones	Desconocido	Mapa de calor
Social Newsrooms	Interés general	Temas relevantes	Estadísticas
ScatterBlogs	Opiniones	Eventos	Clusters en mapa

Tabla 1: Comparación de herramientas de visualización de tweets

Tras el estudio de los distintos sistemas existentes, se observa que existen herramientas de visualización dedicadas a monitorización de eventos, así como herramientas que muestran clusters y etiquetas sobre mapas. No obstante, se aprecia un vacío en la combinación de estas dos características. Es decir, no se han

desarrollado herramientas visuales para monitorización de eventos que muestren clusters de tweets y etiquetas sobre un mapa. Este vacío supone una oportunidad de trabajo que pueda cubrir necesidades que aún no han sido tratadas.

3. Planteamiento del problema

A lo largo de este apartado se explicará cómo ha decidido plantearse la solución al problema descrito anteriormente en este documento.

3.1. Requisitos

Una vez clarificado el problema - procesar y visualizar de forma eficiente toda la información compartida en Twitter durante un evento excepcional -, el primer paso hacia el desarrollo de la solución consiste en especificar los requisitos a cumplir.

3.1.1. Requisitos funcionales

Los requisitos funcionales son aquellos que detallan el "qué" debe realizar el software. La herramienta propuesta debe cumplir con los siguientes requisitos:

Identificador: RE-FU-01	
Nombre: Mapa de la zona afectada	
Prioridad: Alta	**Necesidad**: Alta
Descripción: La herramienta debe cargar un mapa de la zona afectada por el evento que se esté analizando.	

Tabla 2: Requisito RE-FU-01

Identificador: RE-FU-02	
Nombre: Permitir Navegación	
Prioridad: Alta	**Necesidad**: Alta
Descripción: Debe permitir la navegación por el mapa del mundo.	

Tabla 3: Requisito RE-FU-02

Identificador: RE-FU-03	
Nombre: Permitir zoom	
Prioridad: Alta	**Necesidad**: Alta
Descripción: Debe permitir cambiar el zoom, tanto aumentándolo como incrementándolo, adaptando la visualización de en función del zoom.	

Tabla 4: Requisito RE-FU-03

Identificador: RE-FU-04	
Nombre: Mostrar marcadores	
Prioridad: Alta	**Necesidad**: Alta
Descripción: Por cada tweet almacenado en la base de datos, cuyas coordenadas no estén vacías, se debe mostrar un marcador en las coordenadasde dicho tweet.	

Tabla 5: Requisito RE-FU-04

Identificador: RE-FU-05	
Nombre: Agrupar en clusters	
Prioridad: Alta	**Necesidad**: Alta
Descripción: Los marcadores se deben agrupar en clusters, es decir, grupos de marcadores con coordenadas próximas	

Tabla 6: Requisito RE-FU-05

Identificador: RE-FU-06	
Nombre: Visualizar contenido del tweet	
Prioridad: Alta	**Necesidad**: Alta
Descripción: Al hacer click sobre cada marcador, se muestra el contenido del tweet al que corresponda dicho marcador.	

Tabla 7: Requisito RE-FU-06

Identificador: RE-FU-07	
Nombre: Clusters representados por polígonos	
Prioridad: Alta	**Necesidad**: Media
Descripción: Los clusters se representan con un polígono.	

Tabla 8: Requisito RE-FU-07

Identificador: RE-FU-08	
Nombre: Polígonos de área óptima	
Prioridad: Alta	**Necesidad**: Media
Descripción: Los polígonos que representan los clusters, deben ser de tal forma que el área del polígono cubre todos los puntos del cluster de forma óptima, i.e. con el área mínima posible.	

Tabla 9: Requisito RE-FU-08

Identificador: RE-FU-09	
Nombre: Mostrar los términos más frecuentes	
Prioridad: Alta	**Necesidad**: Alta
Descripción: Por cada cluster, la herramienta debe mostrar el término más frecuente entre los tweets pertenecientes al cluster en cuestión.	

Tabla 10: Requisito RE-FU-09

Identificador: RE-FU-10	
Nombre: Análisis Sintáctico	
Prioridad: Alta	**Necesidad**: Alta
Descripción: Los tweets deben ser analizados semánticamente.	

Tabla 11: Requisito RE-FU-10

Identificador: RE-FU-11	
Nombre: Almacenar sustantivos en base de datos	
Prioridad: Alta	**Necesidad**: Alta
Descripción: Del análisis sintáctico de los tweets, se extraen los sustantivos, que se almacenan en la base de datos junto con el id del tweet en el que aparecen.	

Tabla 12: Requisito RE-FU-11

Identificador: RE-FU-12	
Nombre: Clusters de distintos colores	
Prioridad: Media	**Necesidad**: Media
Descripción: Cada cluster se muestra de un color distinto en función de su término más frecuente.	

Tabla 13: Requisito RE-FU-12

Identificador: RE-FU-13	
Nombre: Distintas categorías de términos	
Prioridad: Media	**Necesidad**: Media
Descripción: Existirán distintas categorías de términos, de modo que cada término pertenezca necesariamente a alguna de estas categorías	

Tabla 14: Requisito RE-FU-13

Identificador: RE-FU-14	
Nombre: Un color para cada categoría	
Prioridad: Media	**Necesidad**: Media
Descripción: A cada categoría se le asigna un color. El color en el que se muestre el cluster será el que se haya asignado a la categoría del término más frecuente entre los tweets pertenecientes al cluster.	

Tabla 15: Requisito RE-FU-14

Identificador: RE-FU-14	
Nombre: Añadir animación	
Prioridad: Media	**Necesidad**: Media
Descripción: Los tweets se añaden uno por uno en el orden en el que fueron publicados	

Tabla 16: Requisito RE-FU-15

Identificador: RE-FU-16	
Nombre: Mostrar hora del último tweet añadido	
Prioridad: Baja	**Necesidad**: Media
Descripción: Se muestra la hora a la que se publicó el último tweet que haya sido añadido a la visualización	

Tabla 17: Requisito RE-FU-16

Identificador: RE-FU-17	
Nombre: Recalcular clusters al hacer zoom	
Prioridad: Baja	**Necesidad**: Media
Descripción: Cada vez que se realice zoom, se borran todos los clusters y etiquetas mostrados, y se vuelven a calcular y visualizar los clusters, polígonos y términos más frecuentes.	

Tabla 18: Requisito RE-FU-17

Identificador: RE-FU-18	
Nombre: Leyenda de colores	
Prioridad: Media	**Necesidad**: Alta
Descripción: Se debe mostrar una leyenda con los colores empleados y la categoría de términos a la que corresponde cada color.	

Tabla 19: Requisito RE-FU-18

3.1.2. Requisitos no funcionales

Los requisitos no funcionales definen las restricciones a tener en cuenta a la hora de desarrollar la solución. Los requisitos recogidos a continuación son con los que deberá contar la herramienta propuesta.

Identificador: RE-NF-01	
Nombre: Herramienta en PHP	
Prioridad: Alta:	**Necesidad**: Baja
Descripción: La herramienta debe ser desarrollada sobre PHP.	

Tabla 20: Requisito RE-NF-01

Identificador: RE-NF-02	
Nombre: API de Google Maps	
Prioridad: Alta	**Necesidad**: Baja
Descripción: Se utiliza la API de Google Maps para la visualización del mapa.	

Tabla 21: Requisito RE-NF-02

Identificador: RE-NF-03	
Nombre: Tweets en Base de Datos	
Prioridad: Alta	**Necesidad**: Alta
Descripción: Los tweets se almacenan en una base de datos	

Tabla 22: Requisito RE-NF-03

Identificador: RE-NF-04	
Nombre: Método de Graham	
Prioridad: Media	**Necesidad**: Baja
Descripción: El polígono de área óptima que represente cada cluster se calcula utilizando el método de Graham.	

Tabla 23: Requisito RE-NF-04

Identificador: RE-NF-05	
Nombre: Diseño adaptativo	
Prioridad: Media	**Necesidad**: Alta
Descripción: El diseño de la herramienta debe ser capaz de adaptarse a distintos tamaños de ventanas y dispositivos.	

Tabla 24: Requisito RE-NF-05

3.2. Restricciones y marco regulador

Esta herramienta emplea APIs desarrolladas por terceros, por lo que debe respetar los términos y condiciones de usos de estas.

En primer lugar, recupera la información empleado la API de búsqueda de Twitter [27]. Para ello, se debe aceptar el Acuerdo de Desarrollador de Twitter. En él, el desarrollador se compromete a:

- No realizar ningún proceso de ingeniería inversa sobre la API
- No sobrepasar - ni buscar la forma de sobrepasar - la tasa límite de velocidad de Twitter.
- No utilizar la marca Twitter ni su logo como propia, o insinuar una falsa asociación, patrocinio o aprobación por parte de Twitter.
- No vender los datos recuperados, que pertenecen íntegramente a Twitter
- Proteger al usuario: No distribuir información del usuario que pueda suponerle algún problema legal o pueda potencialmente incumplir la idea de privacidad del usuario
- Mantener la integridad del Twitter: Mostrar Tweets reales, de cuentas reales y sin alteraciones.
- No utilizar la API para generar SPAM
- Evitar replicar el núcleo de la funcionalidad de Twitter
- Exhibir el logo de Twitter.
- No usar contenido recuperado de Twitter para promocionar cualquier producto o servicio sin la autorización expresa del usuario al que pertenece este contenido.
- Incluir botones o iconos de otras redes sociales.
- Usar Tweets de muestra que no existen en la plataforma.

Además, Twitter exige una serie de requisitos para la exhibición de tweets, según los cuales se exige adaptarse a la anatomía de un tweet (Figura 2), realizando alteraciones mínimas.

Figura 2: Anatomía de un tweet

En el caso de la aplicación propuesta, no se han mostrado las acciones que es posible realizar con un tweet, puesto que en este caso la importancia esta en el contenido del tweet, y no en la interacción con él.

Los términos de uso de Twitter hacen referencia también a la Declaración Universal de los Derechos Humanos de la ONU [28]. No se permite mostrar, distribuir, compartir o hacer accesible de modo alguno información a terceros que se considere que puedan emplear esta información para violar de alguna manera esta declaración.

Por otro lado, al aceptar los Términos de Servicio de la API de Google Maps, se acepta:

- Ser el único responsable del uso que hagas de la API
- Estar de acuerdo con que sea Google quien tenga todos los derechos legales sobre el servicio y su contenido.
- Certificas saber y estar de acuerdo con que en ocasiones los servicios sean ofrecidos por terceros.
- No cobrar a los usuarios por utilizar el servicio.
- No copiar, traducir, modificar, replicar el contenido de la API o parte de él.
- No utilizar la API para crear servicios de navegación asistida.

Debe tenerse en cuenta además la **Ley Orgánica de Protección de Datos de Carácter Personal** (LOPD) [29]. Los datos que la herramienta propuesta recoge de Twitter, son datos *abiertos*. Es decir, cualquier persona - registrada o no - tiene acceso a ellos entrando en la plataforma. La Agencia Española de Protección de Datos (AEPD) no identifica a internet como un medio de comunicación y, por tanto, no corresponde a ninguna de las categorías de fuente de acceso público según se definen en la LOPD [30]. Esto significa que es legal re-elaborar y mostrar de diferentes formas la información, pero no darle un uso distinto salvo que se tenga el consentimiento directo de cada usuario.

4. Diseño de la solución técnica

En los siguientes puntos se describirá cómo se ha diseñado la solución técnica al problema descrito en apartados anteriores del presente documento.

Siguiendo los objetivos a cumplir y los requisitos especificados en el apartado anterior, se procede a describir cómo se ha recuperado la información de Twitter, diseñado la base de datos, analizado semánticamente los datos recuperados, categorizado los temas mas relevantes, visualizado en el mapa y finalmente definido los clusters.

4.1. Recuperación de la información

Para extraer la información presente en Twitter, se ha empleado la **API de búsqueda** de Twitter, que permite buscar información entre tweets recientes, i.e. los tweets publicados en los 7 días anteriores al momento de la búsqueda. Para realizar las búsquedas permite emplear diferentes parámetros que permiten especificar con mayor detalle los resultados que se desea obtener.

Es importante destacar que la API que ofrece Twitter para búsquedas se centra en la relevancia en lugar de la completitud. Esto significa que te devuelve únicamente aquellos tweets que Twitter considera de mayor interés y no todos los que realmente se han publicado.

Para tener acceso a esta API, en primer lugar es necesario registrarse como desarrollador en Twitter, con lo que se obtienen las credenciales necesarias para realizar las búsquedas de información.

Se ha empleado un código escrito en PHP para acceder al API y recuperar los datos. En primer lugar, se encarga de crear las tablas de la base de datos en las que se almacenará la información.

A continuación, realiza una conexión con Twitter utilizando los credenciales de desarrollador. Establecida la conexión, se realiza una consulta que busca aquellos tweets en los que aparezca un término deseado. Este término dependerá del caso de uso, i.e. del evento que desee analizar en ese momento. La consulta devuelve una serie de *statuses* que se corresponden con los tweets, organizados en páginas. Por tanto, es necesario recorrer todas las páginas que se han obtenido, y dentro de cada página, todos los statuses. De cada status, se recoge toda la información que se desea almacenar en la base de datos y se inserta con sentencias SQL.

4.2. Diseño de la base de datos

La base de datos se ha decidido implementar utilizando MySQL. Esta base de datos se ha almacenado en un servidor, que durante la realización de este Trabajo de Fin de Grado ha sido local. De esta forma, se permite una conexión entre la base de datos, que almacena la información, y la aplicación PHP encargada de la visualización de cara a los usuarios, permitiendo que en un futuro la base de datos pueda estar en un servidor remoto.

Inicialmente se ha partido de una base de datos que contiene los tweets correspondientes a la consulta que se haya realizado en la fase de recuperación de la información. A continuación, se ha creado un script PHP para conectarse a cada una de ella, extraer la información de los tweets y usuarios contenidos en cada una utilizando consultas SQL, y se han insertado en una nueva base de datos cuyo contenido engloba al de las tres iniciales, tras comprobar antes de cada inserción que la información que se está tratando insertar no está recogida ya en la base de datos.

Puesto que los datos se han recogido utilizando la API de Twitter, la estructura de la base de datos es la que viene dada por defecto por Twitter. Esta estructura se recoge en la siguiente imagen, que muestra el diagrama entidad-relación extendido, o ERR (Enhanced entity–relationship model) de la base de datos (Figura 3):

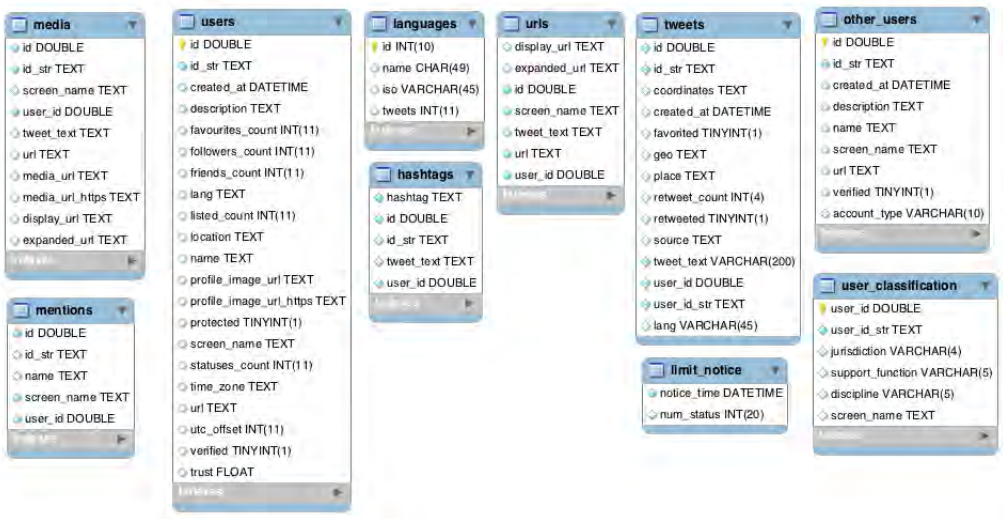


Figura 3: EER de la base de datos original

No obstante, existen algunas tablas que, en el caso concreto de la herramienta propuesta en este trabajo, no son necesarias. Únicamente se desea almacenar la información de los tweets, los usuarios, y los idiomas. Por esto, eliminar el resto de tablas de la base de datos no supondría ninguna pérdida para la aplicación, por lo que se ha tomado de decisión de suprimirlas.

Además, es necesario recoger la información acerca de los términos de cada

tweet, ya que estos son de vital importancia para esta aplicación. Cada termino será analizado semánticamente y representará los tópicos relevantes para los datos a visualizar, lo que implica que se debe crear una nueva tabla en la base de datos para esta información.

Sobre los términos se debe almacenar el término mismo, el tweet en el que aparece, la categoría a la que pertenece y el color con el que se visualizará la categoría. Aunque el color de una categoría y su nombre pueden derivarse el uno del otro, se ha decidido almacenar los dos valores en la base de datos, ya que almacenar el código de color simplifica los cálculos durante la ejecución del programa, y el nombre de la categoría facilita la lectura y verificación de la categoría asignada a cada término.

Por tanto, la estructura de esa tabla será:

TERMINOS (id, termino, tweet, categoría, color)

Finalmente, tras realizar estas modificaciones necesarias sobre la estructura de la base de datos original, el diagrama EER de la base de datos resultante es el siguiente (Figura 4):

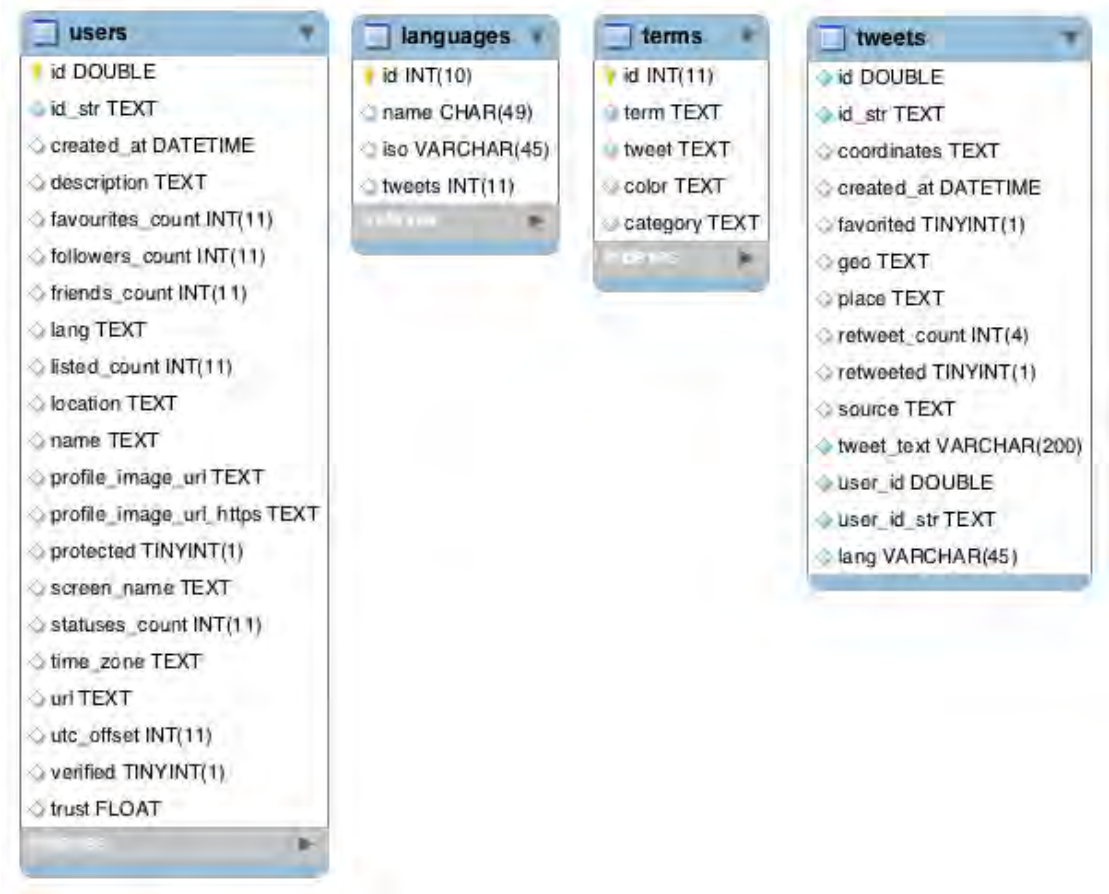


Figura 4: EER de la base de datos final

De los usuarios se almacena la siguiente información:

- `id`: id de la tupla en formato numérico.
- `id_str`: id de la tupla en formato de texto.
- `created_at`: fecha y hora de cuándo se creo el usuario.
- `description`: descripción del usuario.
- `favourites_count`: número de tweets que el usuario ha marcado como favoritos
- `followers_count`: número de usuarios que siguen al usuario al que idenifica la tupla
- `friends_count`: número de usuarios que siguen al usuario al que idenifica la tupla, y a los que el usuario sigue también.
- `lang`: idioma del usuario.
- `listed_count`: número de listas en las que aparece el usuario.
- `location`: ubicación del usuario.
- `name`: nombre completo del usuario.
- `profile_image_url`: URL de la imagen de perfil del usuario, usando el protocolo HTTP.
- `profile_image_url_https`: URL de la imagen de perfil del usuario, usando el protocolo HTTPS.
- `protected`: indica si se trata de un perfil público o privado.
- `screen_name`: nombre de usuario.
- `statuses_count`: número de tweets que el usuario ha publicado.
- `time_zone`: zona horaria en la que se ubica el usuario.
- `url`: página web especificada por el usuario asociada a su perfil.
- `verified`: indica si se trata de una cuenta verificada.
- `trust`: indica si se trata de una cuenta verificada.

De los tweets se almacena la siguiente información:

- `id`: id de la tupla en formato numérico.
- `id_str`: id de la tupla en formato de texto.

- `coordinates`: coordenadas en las que se publicó el tweet.
- `created_at`: fecha y hora de cuándo se creo el usuario.
- `favorited`: indica si el tweet ha sido marcado como favorito.
- `place`: nombre de la ubicación en la que se publicó el tweet.
- `retweets_count`: número de veces que el tweet ha sido retweeteado.
- `retweeted`: indica si el tweet ha sido retweeteado.
- `source`: permite compartir del tweet con formato HTML
- `tweet_text`: contenido del tweet
- `user_id`: id en formato numérico asociado al usuario que escribió el tweet.
- `user_id_str`: id en formato de texto asociado al usuario que escribió el tweet.
- `lang`: idioma en que está escrito el tweet.

4.3. Análisis Sintáctico

Para comprender el contenido de cada tweet que se ha recopilado, ha sido necesario un análisis sintáctico del contenido de estos tweets. Para ello se ha utilizado un etiquetador morfo-sintáctico, o Part of Speech (POS) tagger, cuya función es asignar a cada término su categoría gramatical correcta (Figura 5).

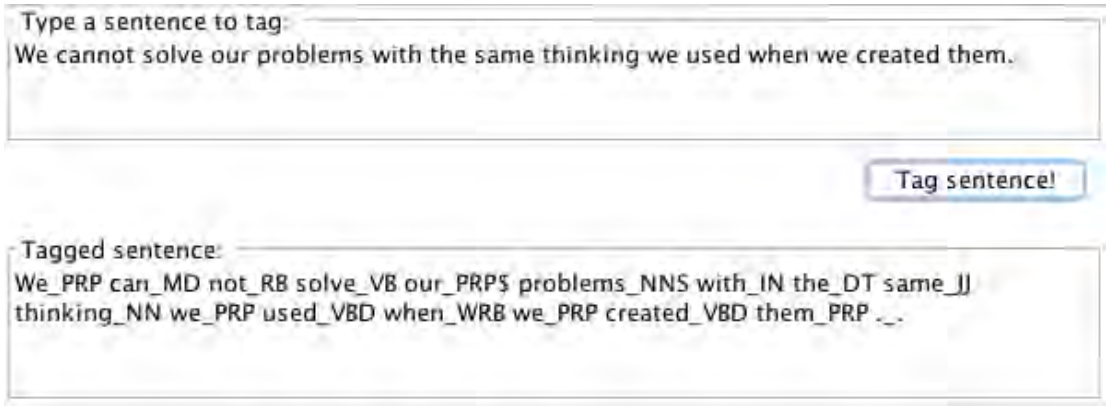


Figura 5: Ejemplo de etiquetador morfosintáctico

Para el desarrollo de esta herramienta, se ha optado por emplear el etiquetador de Stanford: Stanford POS tagger [31]. Esta decisión se ha tomado ya que este etiquetador permite el análisis en distintos idiomas, lo que resulta especialmente útil para analizar eventos ocurridos en otros países o en países con distintos idiomas. Así, se podrían recoger y analizar, por ejemplo, tweets en los distintos idiomas del país, o en el idioma del país y en el idioma de los operadores de los

servicios de emergencia. Además, el etiquetador ofrece también la posibilidad de lematización.

Para realizar este análisis, en primer lugar ha sido necesario extraer los tweets de la base de datos. Esto se ha realizado con un script PHP que, con una consulta SQL, obtiene la colección de tweets para los que se tienen coordenadas. Una vez se tienen estos tweets, el programa genera un fichero con los tweets.

En los casos en los que se decide analizar tweets en varios idiomas, es necesario crear ficheros diferenciados para cada idioma. Esto se debe a que cada idioma es analizado empleado diccionarios distintos, y por tanto es imprescindible separarlos y poder posteriormente indicarle al POS Tagger qué modelo - que depende del idioma - debe utilizar para cada uno.

No obstante, antes de realizar el análisis sintáctico, se ha limpia el contenido de los tweets, eliminando caracteres extraños (puntos, comas, comillas...), saltos de línea y números.

Por cada tweet de la base de datos, se ha comprobado su idioma empleando el valor del campo `lang` (Figura 4). Junto con el contenido - ya limpiado - de cada tweet, se almacena el id del tweet. En el fichero correspondiente - de acuerdo al idioma de los tweets - se almacenan el id y el contenido del tweet. Para identificar cada frase, se escribe un punto tras el contenido de cada tweet.

Generados los archivos de texto con los identificadores y contenido de los tweets, se realiza el análisis sintáctico. Por cada idioma que se desee analizar, debe existir un archivo diferenciado. Estos se analizaran - por separado - empleando el Stanford POS tagger, cuyo resultado se almacena en distintos documentos xml: de nuevo, uno para cada idioma.

El archivo resultante está separado por frases, que a su vez se descomponen en palabras. De cada palabra, nos proporciona la siguiente información:

- `id`: identificador del término dentro de la frase.
- `pos`: categoría gramatical a la que pertenece el término.
- `lemma`: raíz de la palabra.

Se ha procedido a continuación, a la inserción de los términos en la base de datos. Se ha utilizado el campo de POS - o categoría gramatical - para seleccionar únicamente a los sustantivos, ya que se considera que estos son los que mayor significado aportan. También se ha decidido despreciar las palabras con menos de dos caracteres, ya que estas palabras aportan poca información al sistema.

Además se ha realizado un proceso de lematización, de modo que en lugar de almacenar el término en la base de datos tal cual aparece en los tweets, se ha almacenado su raíz gramatical. Con esto eliminamos, por ejemplo, problemas

de plural-singular, de modo que *attack* (ataque) y *attacks* (ataques) serán considerados la misma palabra para la herramienta.

Para la inserción, se han utilizado consultas SQL desde PHP. Gracias a que anteriormente hemos almacenado el id del tweet junto a su contenido, en este paso se ha podido utilizar ese valor para almacenar en qué tweet aparece cada término.

Cabe destacar además, que las combinaciones (TERMINO,TWEET) son únicas. Es decir, no se permite la inserción de un mismo término varias veces para un mismo tweet.

4.4. Categorización de los términos

Realizado el análisis sintáctico de los tweets, se analiza semántica cada término recuperado. Cada uno de estos términos pertenece a una categoría. Se han definido seis posibles categorías:

Categoría	**Descripción**
Emergency	Conceptos propios del dominio de emergencias
Time	Expresiones de tiempo
Place	Listado de países
Media	Vocabulario de medios de comunicación
Evacuation	Conceptos propios del dominio de evacuación
General	Términos que no pertenecen a las categoría anteriores

Tabla 25: Categorías

Para realizar la categorización de términos se ha han utilizado diferentes técnicas de data mining que involucran una ontología y unas taxonomías. La ontología empleada contiene términos de distintos dominios relevantes para la herramienta [25] (Figura 6).

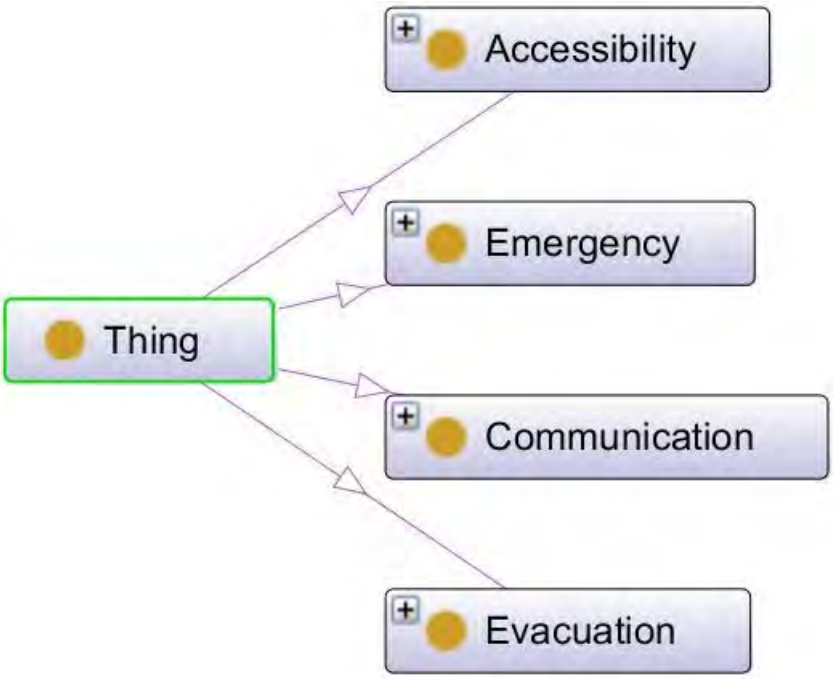


Figura 6: Ontología empleada para la categorización de términos

Con esta ontología se cubren las categorías de *Emergency*, *Media* y *Evacuation*. Aunque también contiene una clase de *Accessibility* para términos relacionados con accesibilidad, en este proyecto no se ha trabajado con ella. No obstante, si en un futuro se decidiera emplearla sería posible incluirla.

Cada una de las cuatro clases principales de la ontología contiene a su vez varias clases y conceptos en una estructura jerárquica. La clase **Emergency** contiene las siguientes sub-clases (Figura 7):

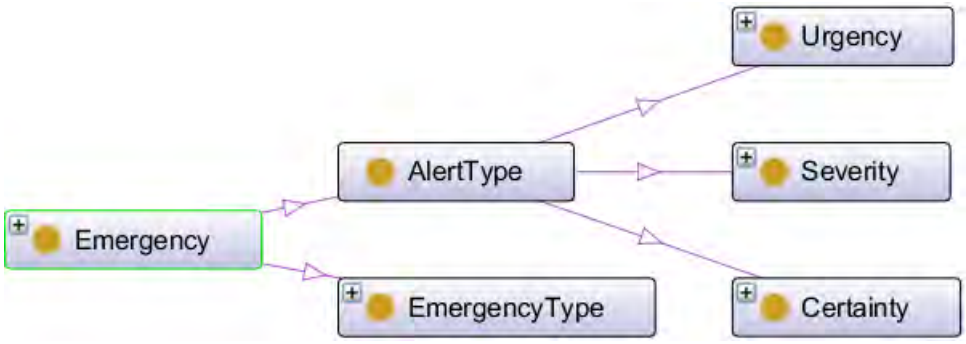


Figura 7: Ontología: Categoría Emergency

Pertenecen a esta clase términos como: *attack* (ataque), *explosion* (explosión) o *tragedy* (tragedía).

La clase **Communication** se organiza de la siguiente manera (Figura 8):

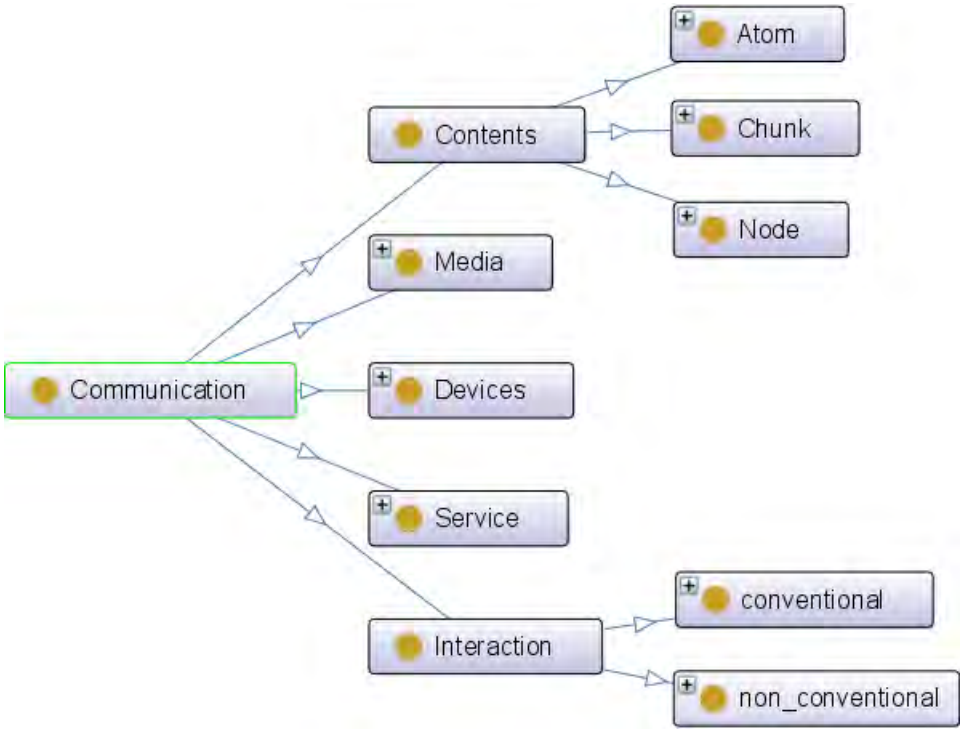


Figura 8: Ontología: Categoría Communication

La clase Communication representa la categoria **Media** y contiene términos como *video*, *sound* (sonido) o *radio*.

Esta categoría es a la que menos términos pertenecen. Este no es un hecho preocupante puesto que también es la categoría - de las recogidas en la ontología - cuyos términos son de menor importancia en el dominio en el que se pretende usar la herramienta.

Por último encontramos la categoría **Evacuation**, que como las demás se descompone en distintas sub-categorías. Su estructura es la siguiente (Figura 9):

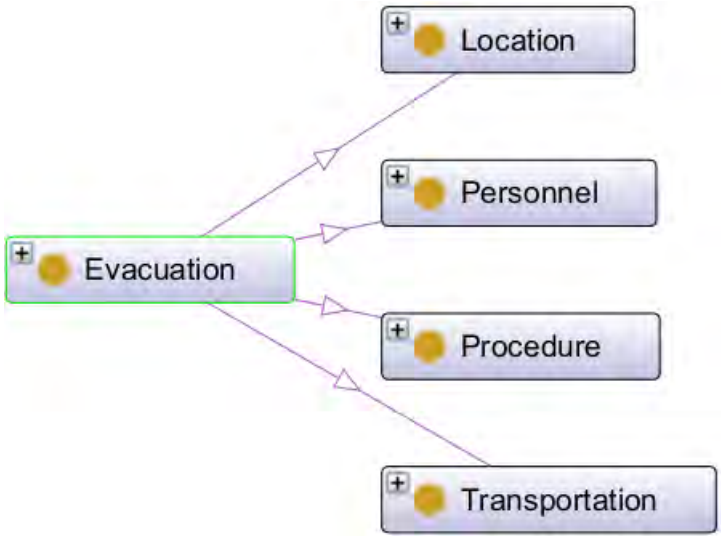


Figura 9: Ontología: Categoría Evacuation

Pertenecen a esta categoría términos como: *shelter* (refugio), *route* (ruta) o *location* (ubicación).

Sin embargo, existen más categorías, que no aparecen en la ontología. Se trata de Place, Time y General. Para determinar los términos que pertenecen a estas categorías se han empleado otros métodos.

En primer lugar, se ha empleado WordNet [32]- una base de datos léxica con términos en ingles, que agrupa en conjuntos de sinónimos - para ver si el término en cuestión era sinónimo de *time* (tiempo) o *place* (lugar). De ser así, su categoría será aquella de la que sea sinónimo.

Si no es sinónimo de ninguno, se ha empleado una taxonomía para comprobar si pertenece a la categoría **Time**. Se trata de una lista con expresiones de tiempo. Si el término se encuentra en la lista, su categoría será Time. Algunos términos de esta categoría son: *evening* (tarde), *tonight* (esta noche) o *yesterday* (ayer).

Se ha empleado otra taxonomía para encontrar los términos que pertenecen a la categoría **Place**, que consiste en una lista con nombres de países. Se han encontrado, por ejemplo, *France* (Francia), *Germany* (Alemania) o *America* (América).

A los términos que no pertenecían a ninguna de las categorías anteriores, se les ha asignado la categoría **General**.

Puesto que existen distintos niveles de categorías en la ontología, los términos pueden pertenecer a cualquiera de ellos. Por tanto, tras realizar este proceso de categorización de términos se ha comprobado de qué categorías se han encontrado términos (Figura **??**):

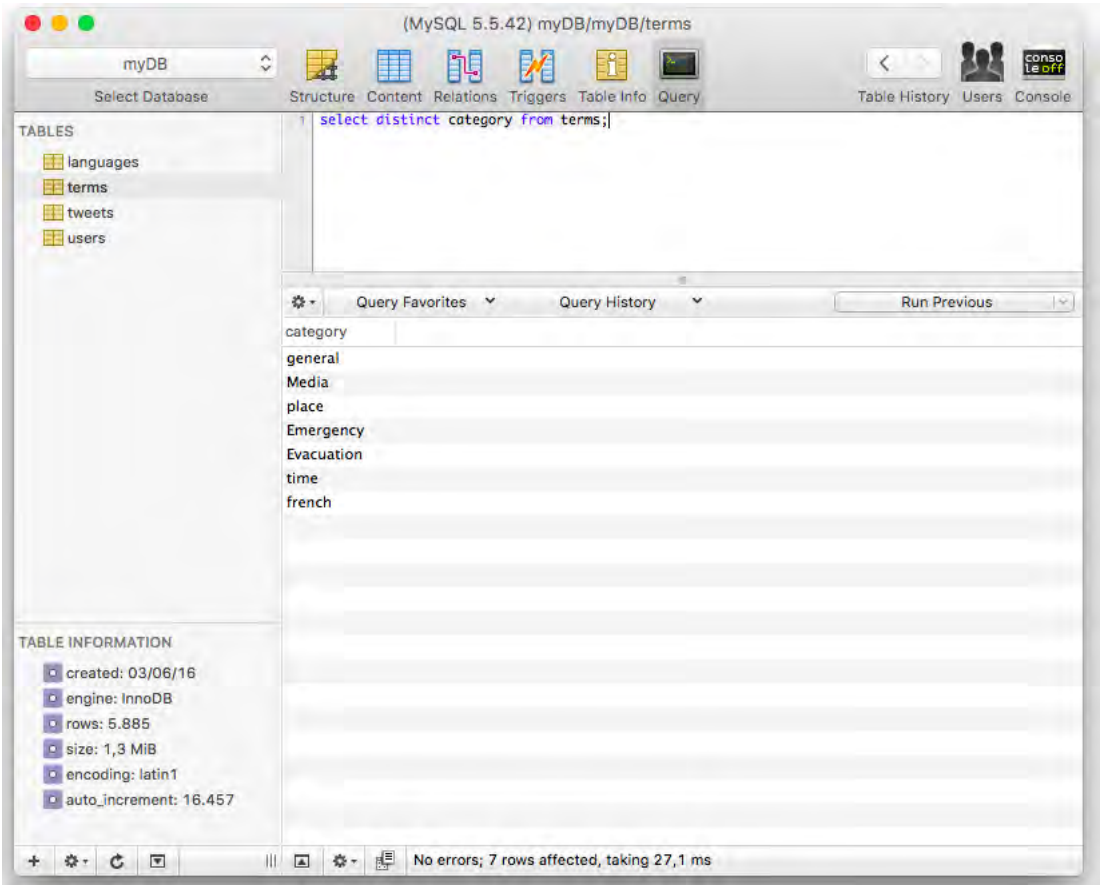


Figura 10: Categorías en las que se han encontrado términos

Por tanto, se ha trabajado únicamente con las categorías en las que hay términos: son a estas a las que se les ha asignado colores para su visualización, y las que aparecen en la leyenda en el mapa de la herramienta.

4.5. Visualización del mapa utilizando la API de Google Maps

Para la visualización del mapa de la zona afecta se ha empleado la API de Google Maps. Esta API te permite mostrar mapas en una página web, y personalizarlos.

Para cargar un mapa es necesario especificar:

- `center`: Coordenadas que representan el centro del mapa.
- `zoom`: Nivel de zoom sobre esas coordenadas.

Existen cuatro tipos básicos de mapas (Figura 11), que se identifican a continuación:

Tipo	Descripción
Roadmap	Visualización por defecto. Mapa básico en 2D.
Satellite	Mapa fotográfico.
Hybrid	Combina los dos anteriores.
Terrain	Mapa geográfico.

Tabla 26: Tipos básicos de mapas

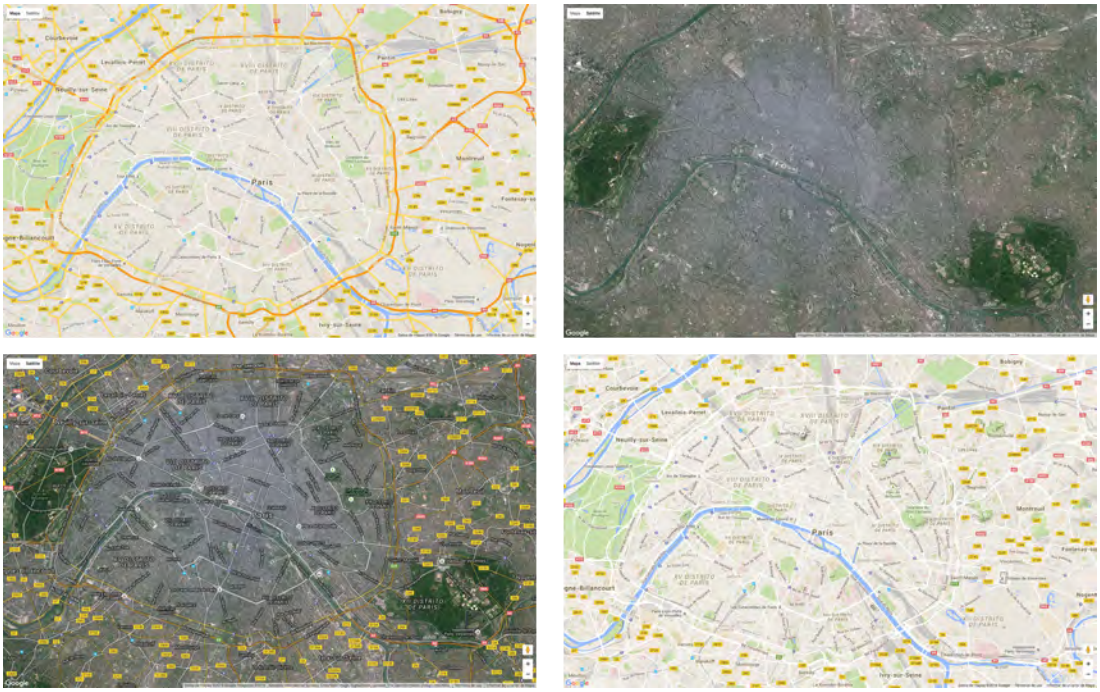


Figura 11: Roadmap (esquina superior izquierda), Satellite (esquina superior derecha), Hybrid (esquina inferior izquierda) y Terrain (esquina inferior derecha)

Se ha elegido emplear mapas de tipo roadmap al ser los más sencillos. La información más importante de la herramienta está en los elementos que se sitúan sobre el mapa. Por tanto, se ha buscado un mapa que permitiese centrar la atención en estos elementos.

Además , la API te permite insertar elementos de gran utilidad para esta herramienta, y , como ya se ha mencionado, ofrece librerías con funcionalidades extra.

Uno de estos elementos, de vital importancia para esta aplicación, son los **marcadores**, que identifican una ubicación en el mapa, para lo que necesitan, al menos, los siguientes parámetros:

- `position`: Latitud y longitud, por ese orden, en las que se desea insertar el marcador.

- `map`: Mapa en el que va a situar el marcador.

En esta herramienta, los marcadores se utilizan para representar las coordenadas en las se ha publicado un tweet guardado en la base de datos.

La imagen que representa a un marcador es personalizable. Puesto que el color identificativo de la plataforma Twitter es el azul, se ha optado por emplear el mismo para representar los tweets. Además, el color rojo - color por defecto de los marcadores - se le ha asignado a una de las categorías de términos, por lo que podía dar lugar a confusiones.

A un marcador, se le puede asociar una **ventana de información**. Se trata de una ventana emergente que muestra contenido sobre el mapa. Esta asociación se realiza con un evento `click` y la función `infowindow.open(map, marker)`, que recibe una ventana y un marcador, de modo que la ventana se abre al hacer click sobre el marcador (Figura 12).

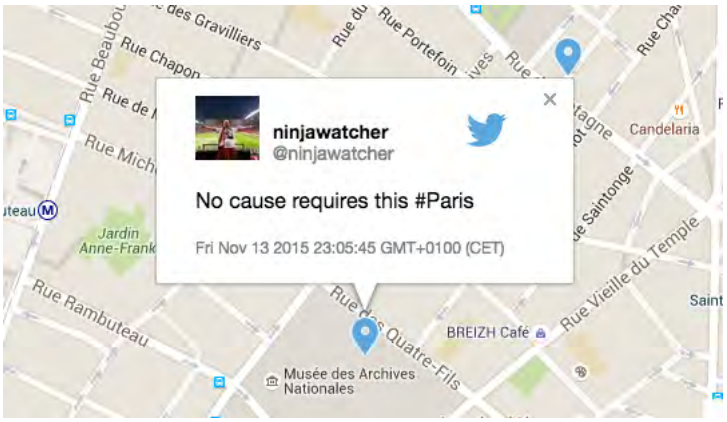


Figura 12: Ventana de información

Se han empleado ventanas de información para mostrar el contenido del tweet al que corresponden las coordenadas especificadas en el marcador.

Una de las librerías que pueden utilizarse con la API de Google Maps, es la librería **markerClusterer**, que permite agrupar los marcadores en clusters en función de su proximidad geográfica. De este modo, se evita sobrecargar el mapa visualmente.

Para apreciar la diferencia, a continuación se muestran dos imágenes en las que se visualizan marcadores utilizando la misma base de datos sin clustering (Figura 13) y con clustering (Figura 14).

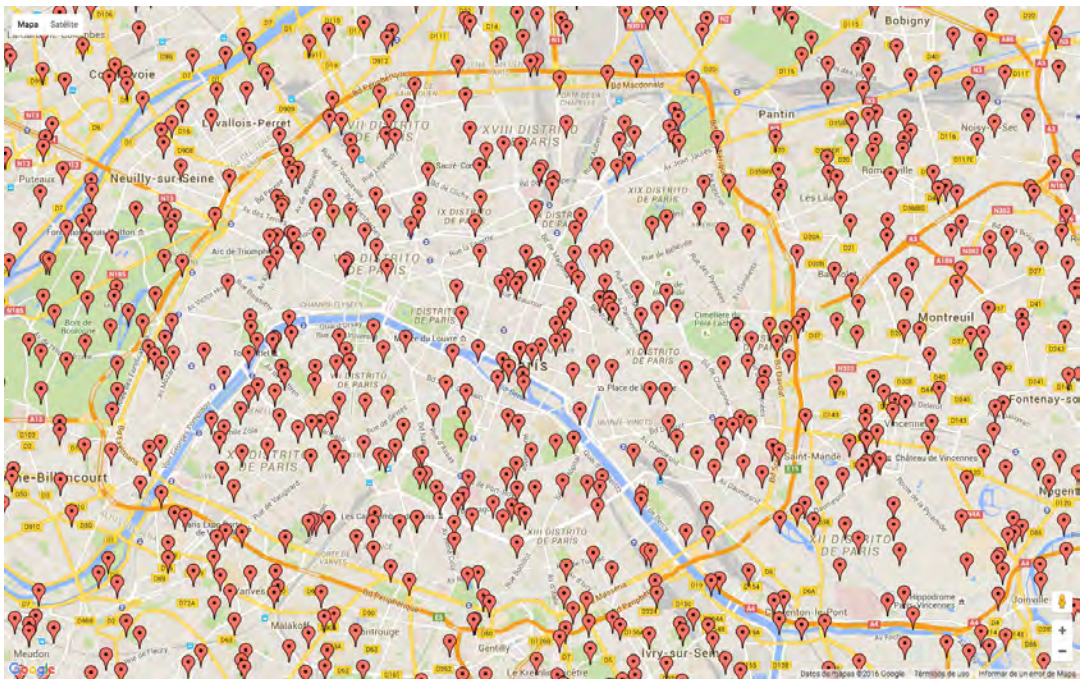


Figura 13: Ejemplo de visualización sin clustering

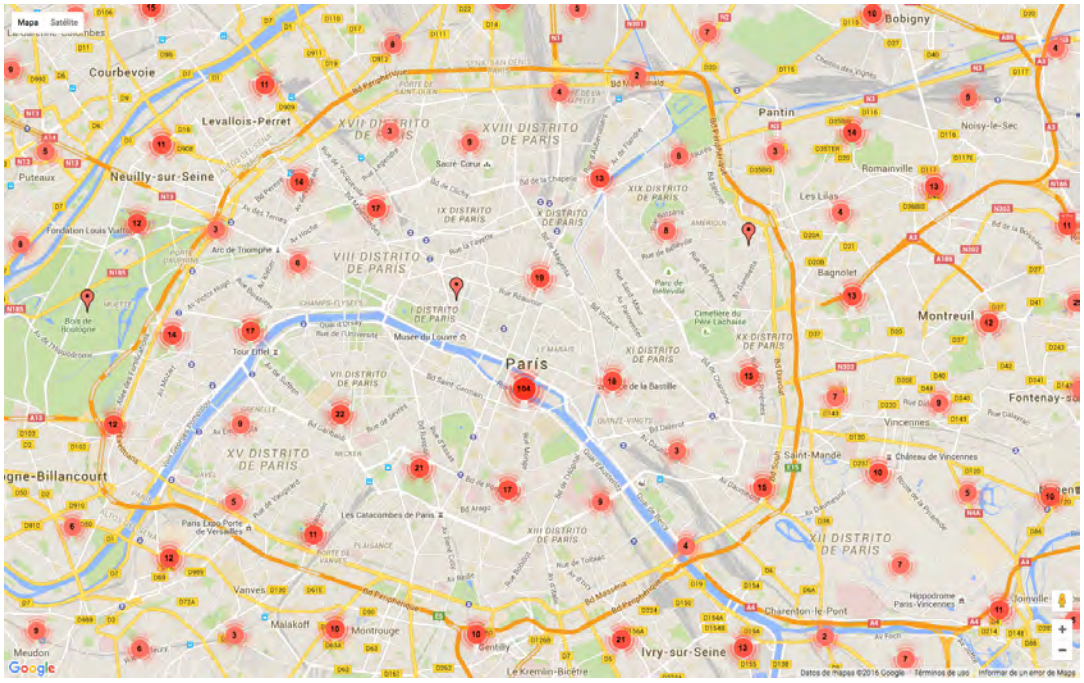


Figura 14: Ejemplo de visualización con clustering con la librería markerclusterer

Existe también la opción de representar **polígonos** sobre el mapa, especificando las coordenadas de los vértices, el color y el grosor de los bordes, y el color y opacidad de relleno (Figura 15).

Figura 15: Polígonos

En esta herramientra, se han empleado polígonos para representar los clusters de marcadores en lugar de utilizar la visualización por defecto de la librería. Esto se explicará detalladamente más adelante.

Una librería que complementa perfectamente a los polígonos, es **maplabel**, que permite añadir etiquetas al mapa. Para ello, se debe especificar su posición, mapa, y texto que se desea mostrar (Figura 16).

Figura 16: Etiquetas

Se han empleado etiquetas para representar el término más frecuente entre los tweets que pertenecen a un cluster.

Además, la API de Google Maps está desarrollada en javascript, por lo que puede ser integrada en una página web. Con todos estos elementos se ha llegado a la siguiente visualización final (Figura 17):

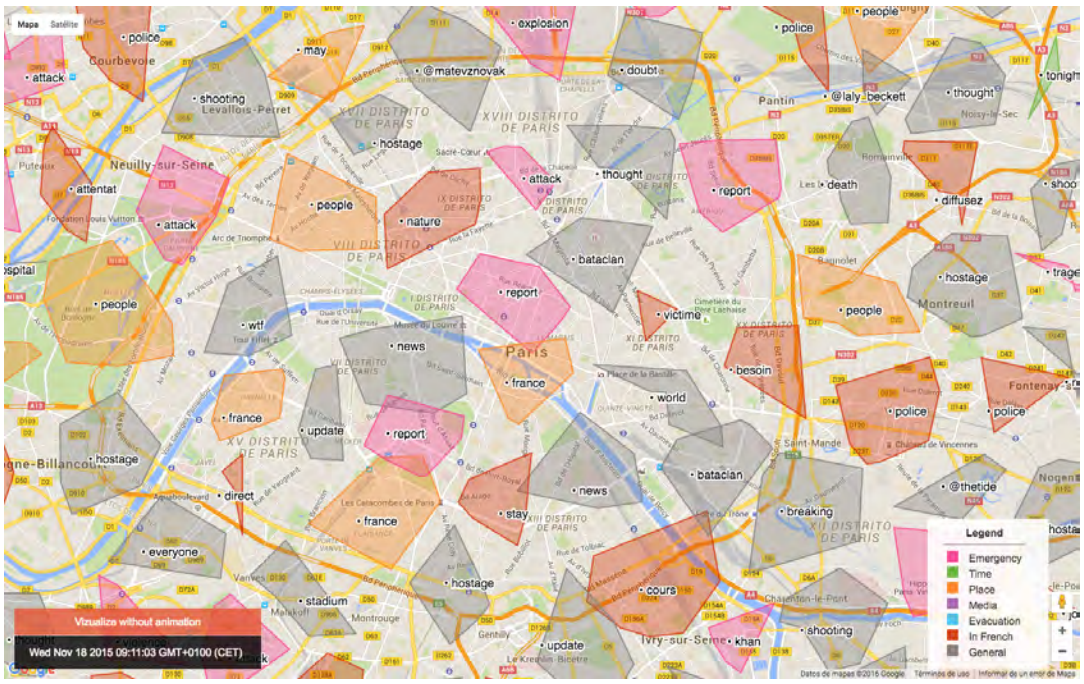


Figura 17: Aspecto final de la herramienta

El color de cada cluster depende de la categorización que se ha hecho:

- **Rosa:** Emergency
- **Verde:** Time
- **Naranja:** Place
- **Morado:** Media
- **Azul:** Evacuation
- **Gris:** General

A modo de animación, se ha implementado la herramienta de modo que en lugar de situar todos los marcadores de una sola vez, se han sitúan de uno en uno.

El cuadro que se observa en la esquina inferior izquierda indica a qué hora se publicó el último tweet que se se ha añadido al mapa.

El objetivo de esta animación es facilitar la comprensión de cuál fue la evolución de la actividad en las redes sociales.

También se puede observar un cuadro en la esquina inferior derecha con una leyenda que indica a qué categoría corresponde cada color de cluster.

4.6. Clustering

El clustering consiste en agrupar una serie de marcadores con el fin de simplicar la visualización y evitar así sobrecargar la pantalla con información.

Aunque se puede agrupar atendiendo a distintos criterios, en esta aplicación se ha decidido agrupar en función de la distancia geográfica entre las coordenadas de los marcadores.

Para realizar el clustering, se ha empleado la librería **markerClusterer** para la API de Google Maps. Esta librería ofrece la opción de personalizar los clustering. Un parámetro especialmente útil es `gridSize`, que determina el tamaño de la región tal que los marcadores de esta región son agrupados dentro del mismo cluster.

En las siguientes imágenes se muestra el aspecto de la herramienta con distintos valores de `gridSize` (Figura 18), partiendo de su valor por defecto (50).

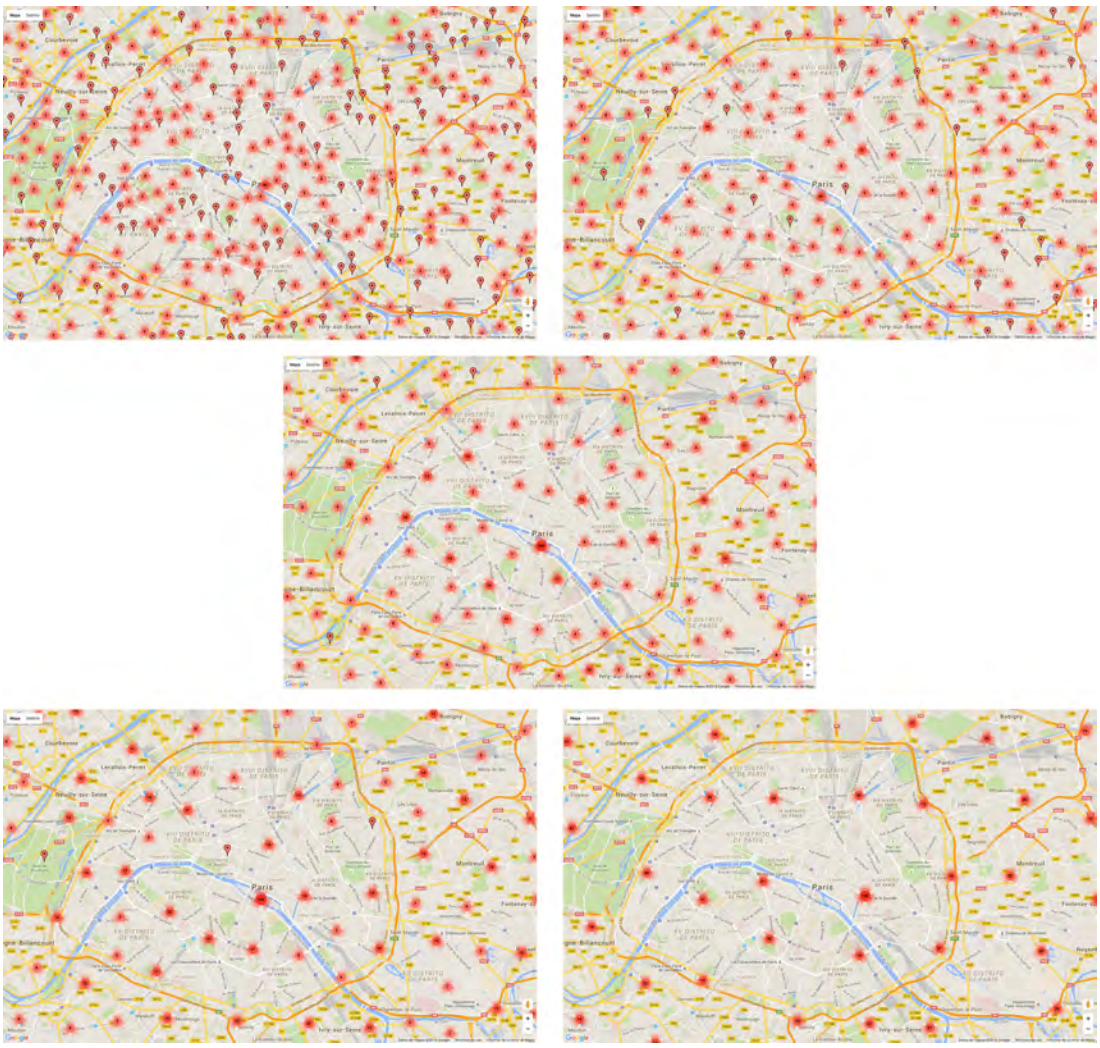


Figura 18: Clustering con un gridSize de valor 30 (esquina superior izquierda), 50 (esquina superior derecha), 75 (centro), 100 (esquina inferior izquierda) y 150 (esquina inferior derecha)

Se ha recogido el número de clusters y el tamaño medio de clusters para cada uno de los valores de `gridSize` que se ha probado, en una herramienta con 2288 marcadores. Los resultados que se han obtenido pueden verse en la siguiente tabla:

gridSize	número de clusters	tamaño medio del cluster (en marcadores)
30	379	2,1266490765171504
50	231	3,675324675324675
75	140	6,478571428571429
100	100	9,53
150	52	20,076923076923077

Tabla 27: Comparación entre distintos valores de gridSize

Al aumentar el tamaño de `gridSize`, disminuye el número de clusters.

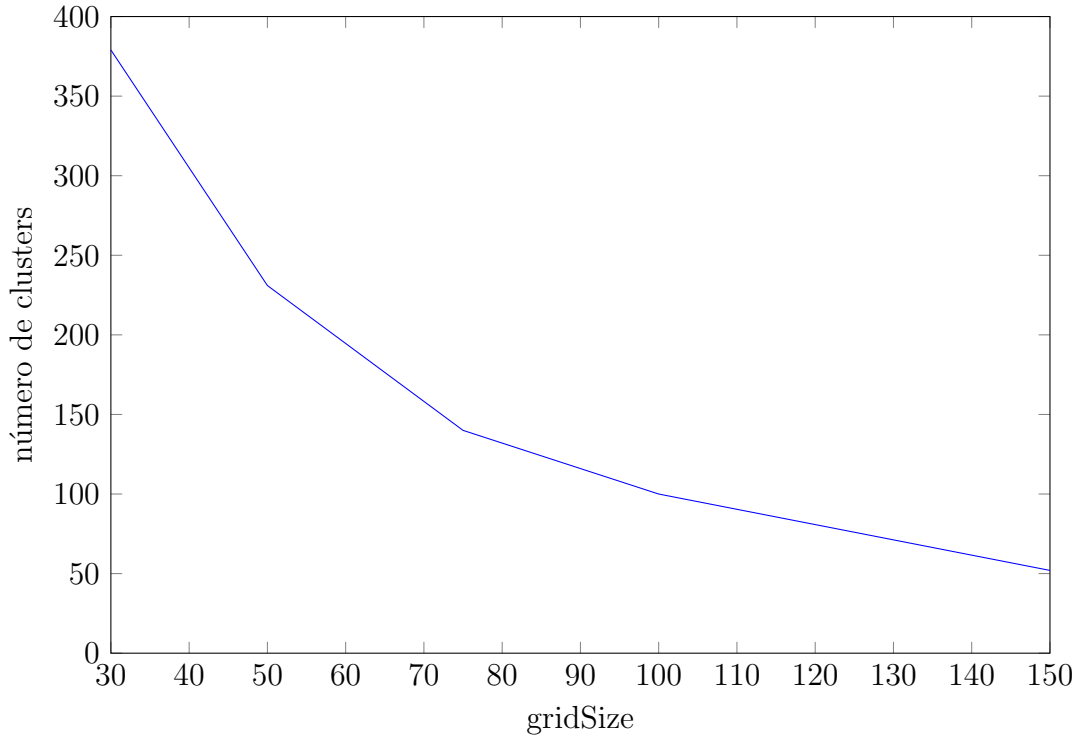


Figura 19: Relación entre el número de clusters y gridSize

No obstante, si el tamaño de `gridSize` aumenta, también lo hace el tamaño de los clusters.

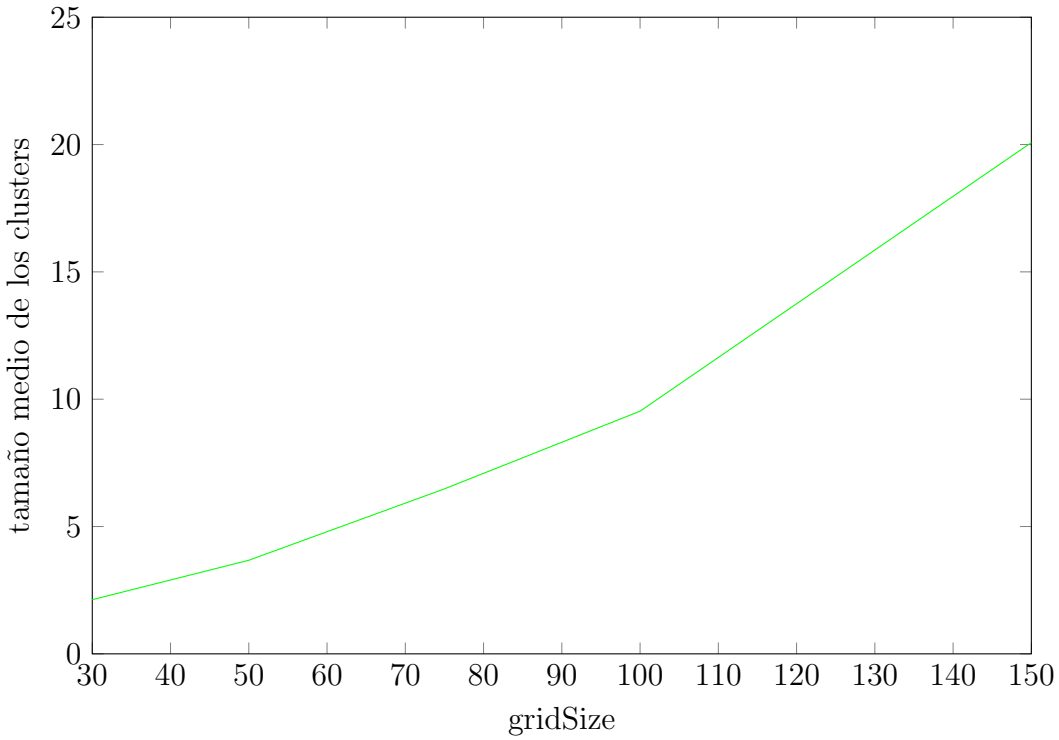


Figura 20: Relación entre el tamaño medio de los clusters y gridSize

Este resultado es el esperado, ya que el número de marcadores es el mismo independientemente del valor de `gridSize`. Por tanto, cuantos más clusters haya, menos marcadores habrá en cada uno. Y cuantos menos clusters haya, más marcadores en cada uno de ellos.

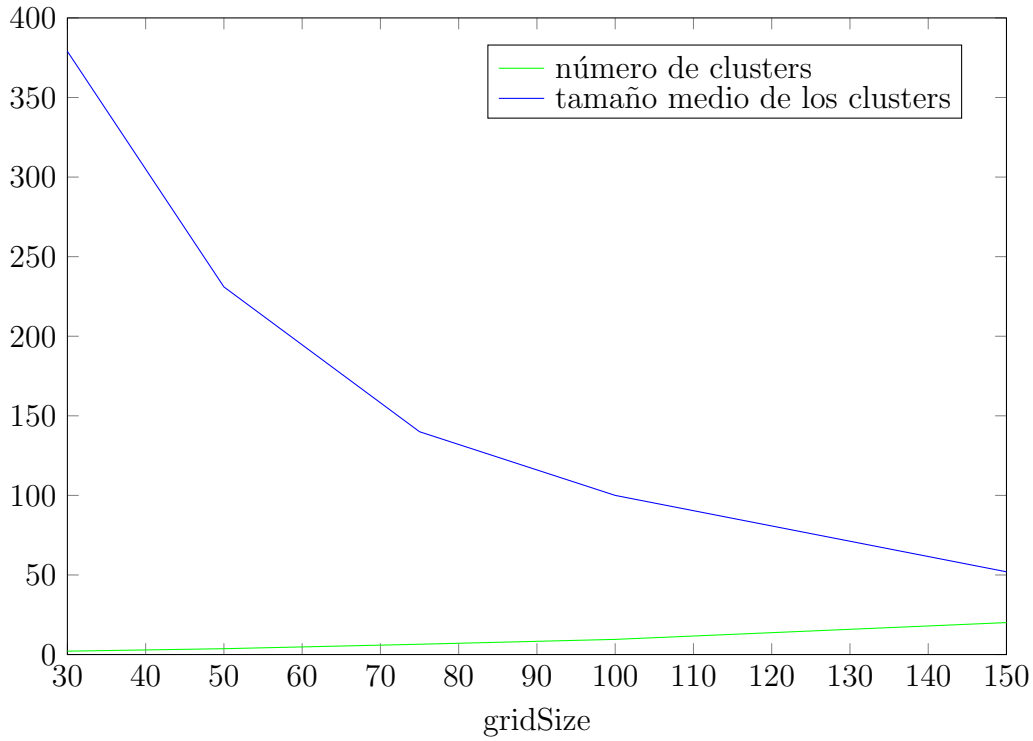


Figura 21: Relación entre el número de clusters y el tamaño medio de los clusters.

Un tamaño medio de cluster muy bajo hace que el resultado se aproxime a no utilizar ningún mecanismo de clustering. Por otro lado, un tamaño demasiado grande, supone una pérdida de información. Se desea encontrar por tanto un término intermedio.

Comparando los resultados obtenidos, y las distintas visualizaciones, se ha optado por usar un tamaño de `gridSize` de 100.

Aunque la funcionalidad de la librería era de gran utilidad para la implementación de la herramienta, se ha optado por otro tipo de visualización más exacta (Figuras 23 en lugar de la visualización por defecto (Figura 22)).

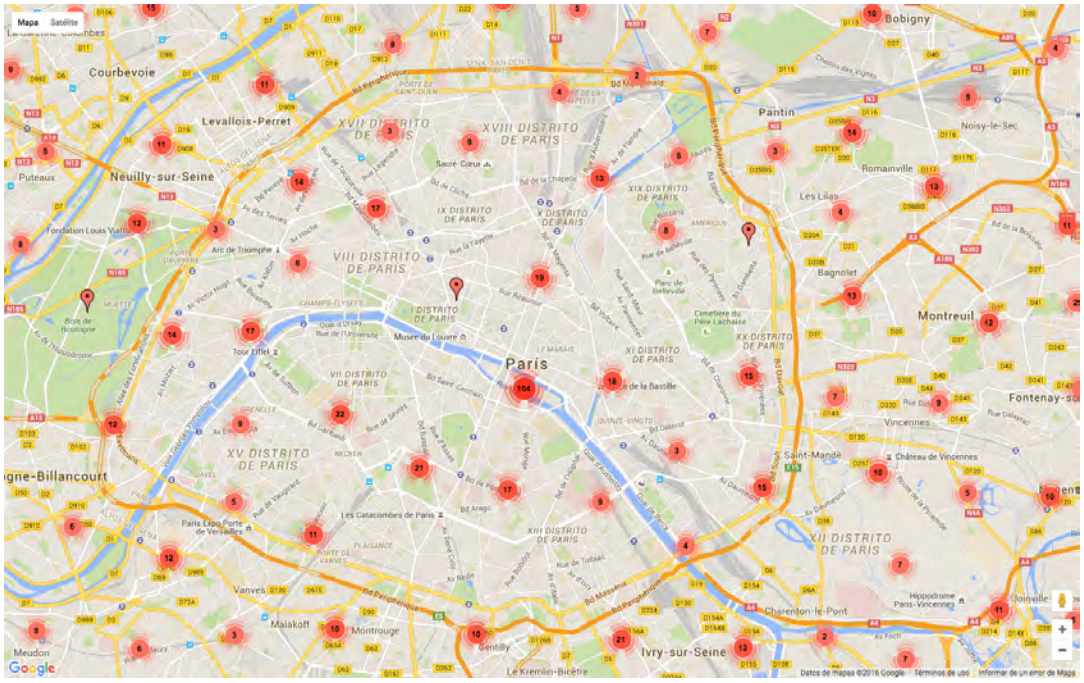


Figura 22: Visualización de clusters por defecto.

Figura 23: Visualización de clusters representados por polígonos.

En esta nueva visualización cada cluster está representado por un polígono, cuya área cubre todos los puntos pertenecientes al cluster.

De esta forma, se puede apreciar en qué zonas exactamente se han publicado los tweets a los que corresponden los marcadores del cluster, a diferencia de la visualización por defecto, que simplemente señala el centro geográfico del cluster.

4.6.1. Método de Graham

El método de Graham, o Graham Scan, es una variación del algoritmo de la envolvente convexa. Dada una serie de puntos, la envolvente convexa es el polígono convexo de menor área que recubre todos los puntos.

Se define como polígono convexo aquel que todos sus ángulos interiores formen ángulos inferiores a 180º y todas sus diagonales son interiores. Es decir, si se traza una línea recta entre dos puntos cualesquiera del polígono, siempre quedará dentro del polígono (Figura 24).

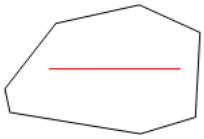


Figura 24: Algoritmo Concave hull: Ejemplo de polígono convexo

A continuación se muestra un polígono no convexo que permite apreciar la diferencia (Figura 25).

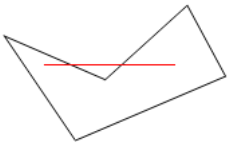


Figura 25: Algoritmo Concave hull: Ejemplo de polígono no convexo

Así, se define como algoritmo de la envolvente convexa aquel que, dada una serie de puntos, calcula su envolvente convexa 26.

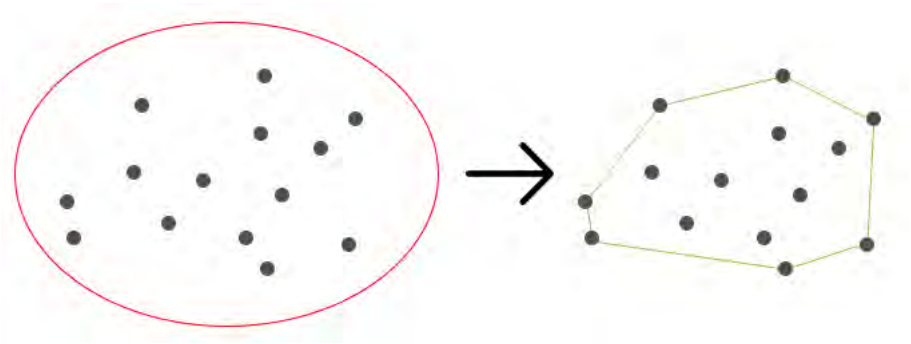


Figura 26: Ejemplo de envolvente convexa

A través de un ejemplo sencillo, se explicará el funcionamiento del Método de Graham partiendo de una serie de puntos (Figura 27).

Figura 27: Método de Graham: serie de puntos inicial

El primer paso será determinar el punto por el que se empieza, llamado P0, que será aquel con el menor valor de Y. En caso de empate, se elegirá aquel con el menor valor de X (Figura 28).

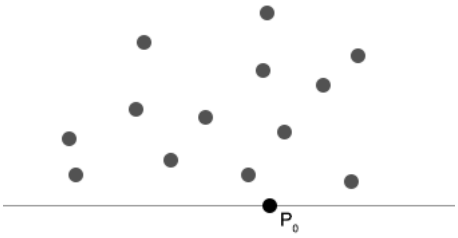


Figura 28: Método de Graham: paso 1

El punto P1 será aquel que unido a P0 forme un mayor ángulo con la recta vertical que cruza P0. Este punto será P1. En los puntos siguientes del algoritmo, no podrá volver a ser considerado como candidato. (Figuras 29 y 30)

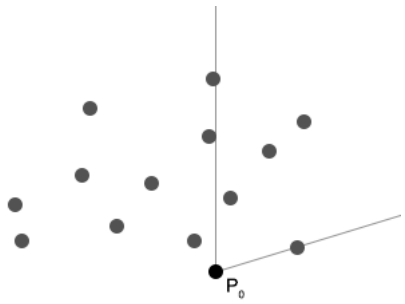


Figura 29: Método de Graham: paso 2a

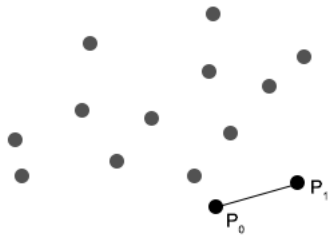


Figura 30: Método de Graham: paso 2b

El punto P2 será aquel cuya recta que lo une con P1 forme un mayor ángulo con la recta que une P1 y P0. Igualmente, no volverá a ser considerado en los pasos siguientes. (Figuras 31 y 32)

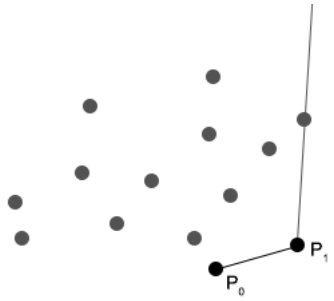


Figura 31: Método de Graham: paso 3a

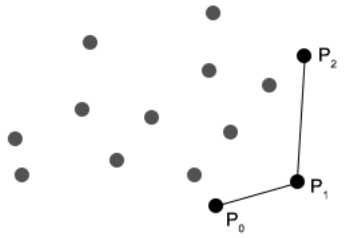


Figura 32: Método de Graham: paso 3b

Se repiten los pasos anteriores hasta que el siguiente punto seleccionado sea el punto por el que se comenzó. (Figura 33)

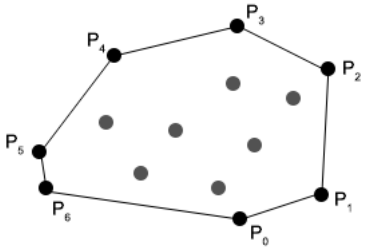


Figura 33: Método de Graham: estado final

Para poder emplear este algoritmo en nuestra herramienta, se ha utilizado una implementación ya realizada del algoritmo en javascript, que se ha adaptado para las coordenadas de Google Maps.

Con dos bucles anidados se han recorrido todos los clusters, y todos los marcadores de cada cluster, de los que se han extraído las coordenadas, que se han almacenado en un array. Se le ha pasado estas coordenadas al método getHull(), que implementa el método de Graham y que devuelve la lista vértices del polígono.

Al recorrer todos los marcadores de cada cluster, no sólo se han almacenado sus coordenadas, sino también sus términos. Así, ahora se comprueba cuál es el término más frecuente de entre los tweets del cluster, y éste determina de qué color se representará el polígono.

5. Caso de estudio

Como caso de estudio durante este Trabajo de Fin de Grado se han utilizado tweets recogidos durante los atentados del 13 de noviembre de 2015 en París.

Se ha elegido este evento por su proximidad tanto geográfica - ya que Francia es un País vecino de España - como temporal - al haberse producido durante este mismo curso académico.

Se trata además de un suceso que tuvo un gran impacto en la sociedad española, puede que precisamente por esta proximidad geográfica, o puede que fuese porque España sufrió también algo similar en 2004 con los atentados en la estación de trenes de Atocha, en Madrid.

El proceso de recolección de datos ha consistido realmente en tres sub-procesos, correspondientes a tres consultas distintas:

- **Paris**: La primera consulta se limitaba a recolectar aquellos tweets en los que se apareciese la palabra *Paris*.
- **Fusillades**: Al ser un evento ocurrido en Francia, se ha querido recoger tweets escritos directamente por la población francesa. Para ello, se ha buscado *fusillades*, que es el término francés para referirse a *fusilamientos*, y que por tanto, fue empleado por algunos usuarios con relación a la tragedia.
- **#PorteOuverte**: Este hashtag, que traducido al castellano significa *Puertas abiertas*, comenzó a utilizarse entre la población francesa o residente en Francia a modo de apoyo. Los ciudadanos ofrecían un lugar en su casa a aquellos que buscasen un lugar seguro en el que refugiarse. (Figuras 34 y 35)

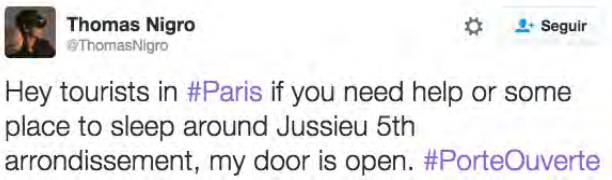


Figura 34: Tweet con el hashtag #PorteOuverte 1

Figura 35: Tweet con el hashtag #PorteOuverte 2

No obstante, la aplicación podrá ser utiliza para cualquier evento que se desee analizar, debiendo modificar únicamente el contenido de la base de datos.

Para este caso de uso, se han encontrado un total de 1.739 términos distintos, siendo los más frecuentes *France* (178 ocurrencias), *attack* (158 ocurrencias) y *people* (147 ocurrencias). Estos datos se han recogido tras eliminar de la base de datos el término Paris junto con los hashtags. Estos términos se han suprimido puesto que eran el término más frecuente de casi la totalidad de los clusters. Ocultarlos permite observar con más detalle la diferencia de términos empleados (Figura 36).

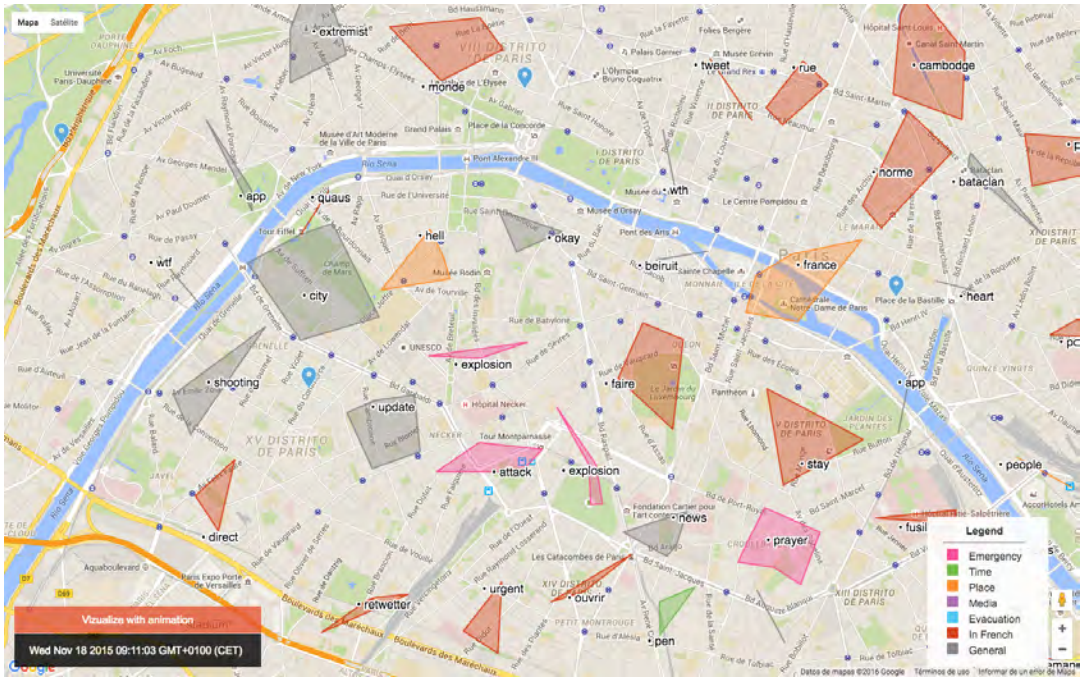


Figura 36: Imagen de la herramienta

5.1. Alternativas de diseño

Durante la implementación de la herramienta, se ha debido tomar una serie de decisiones, en las que se han considerado distintas opciones, eligiendo una de ellas. En este apartado se describen cuáles han sido estas alternativas de diseño y se describe por qué se ha tomado cada decisión.

En primer lugar, se ha decidido recuperar la información de **Twitter**. Los motivos por los que ha decidido así se enumeran a continuación:

- **API de búsqueda**: Twitter proporciona una API de búsqueda que permite recuperar información en función de distintos parámetros. Esto nos permite extraer aquellos tweets escritos que traten un evento excepcional concreto.
- **Popularidad**: Twitter es una de las redes sociales más populares. Esto supone una ventaja puesto que cuanto más popular sea la red social, más información se podrá recoger.

- **Publicaciones en formato de texto**: Mientras que algunas redes sociales se centran en compartir imágenes o vídeos, las publicaciones de Twitter permiten distintos formatos de contenido, lo que permite analizarlo de una forma más sencilla. Además, limita el texto a 140 caracteres, lo que lleva a los usuarios a ser concisos y exactos con lo que escriben.
- **Perfiles públicos**: Cada vez es más frecuente que los usuarios creen perfiles públicos en redes sociales. No obstante, en determinadas redes - por ejemplo, Facebook - los usuarios tienden más a crear perfiles privados, a los que no se tiene acceso sin autorización del usuario, y de los que por tanto no se puede extraer información.

Para mostrar los matas, se ha decidido emplear la **API de Google Maps** para mostrar los mapas. Se ha elegido esta API por distintos motivos:

- **Popularidad**: Los mapas de Google Maps son unos de los más utilizados. Usar un servicio de mapas popular supone una gran ventaja para la herramienta, ya que gran parte de los usuarios probablemente ya estén familiarizados con los mapas, y no necesiten más información sobre como interactuar con ellos.
- **Implementación web**: La API se encuentra desarrollada en javascript, lo que permite incorporarla fácilmente en una página web.
- **Recursos disponibles**: Gracias a la popularidad de Google Maps, se puede encontrar una gran cantidad de librerías con funcionalidades extra desarrolladas para Google Maps. También se puede encontrar documentación muy detallada, y cuenta con una gran comunidad para resolución de dudas.
- **Personalizable**: Los mapas son altamente personalizables. Esto supone un gran número de posibilidades y facilidad de adaptación a lo que se está buscando. Por ejemplo, permite elegir el centro del mapa mostrado inicialmente y el nivel del zoom, con lo que se permite enfocar el mapa en la zona afectada por el evento que se esté tratando.

Para la **visualización de clusters**, se ha empleado una visualización de clusters distinta a la que la librería utiliza por defecto. Se ha optado por una visualización representada por polígonos. Esto tiene la ventaja de que permite visualizar en qué zona exactamente se está publicando el contenido. En cambio, la visualización original únicamente señalaba el centro geográfico del área en el que se estaba compartiendo información.

Desde un principio se ha buscado representar cada terminó en un color diferente con el fin de **simplificar la diferenciación visual de los términos**. No obstante, se han barajado distintas formas de hacerlo. En un principio, se utilizó un color para cada termino. Finalmente, se decidió organizar los términos por categorías. Así, se permite al usuario determinar la importancia de cada término en función

de su categoría. Por ejemplo, para un operario de emergencia, no es lo mismo el término *ayer* que el término *asesinato*. Por tanto, no deberían tratarse de la misma forma.

Se ha debido elegir también entre una visualización con o sin **animación**. Añadir animación permitía comprender la evolución de la actividad a lo largo del tiempo, a la vez que prepara la herramienta para ser implementada en tiempo real. Sin embargo, también implica tener que esperar más en el sentido de que en un principio hay poca información reflejada en la herramienta. Hasta que no ha pasado un tiempo, no se empiezan realmente a distinguir las distintas zonas de actividad - identificadas por clusters - y términos más frecuentes. Como solución, se han implementado los dos modelos de visualización. Así, es el usuario el que elige cómo prefiere que se le presente la información.

5.2. Evaluación

Para evaluar el proyecto, se ha realizado un proceso de evaluación en dos fases.

5.2.1. Fase 1

En la primera fase de evaluación del proyecto, se le ha mostrado la herramienta a una serie de personas ajenas a ella, que han actuado como evaluadores. Sin darles ninguna indicación sobre su uso o finalidad, se le ha pedido que interactúen con ella.

Luego se les han facilitado la siguiente matriz, y se les ha pedido marcar con una X la casilla que considerasen para cada parámetro de evaluación.

	0 (Nada)	1 (Poco)	2 (Regular)	3 (Bastante)	4 (Mucho)
Facilidad de uso:					
Utilidad:					
Interfaz de usuario:					
Objetivo claro:					

Tabla 28: Matriz de evaluación del proyecto para la fase 1

El nombre de los evaluadores ha sido omitido para preservar su confidencialidad, pero a continuación se muestran algunos datos de sus perfiles:

Persona	Profesión	Edad
1	Estudiante	19
2	Óptico	49
3	Informático	50

Tabla 29: Perfil de los evaluadores

Sus evaluaciones han sido las siguientes:

Evaluador	Facilidad de uso	Utilidad	Interfaz de usuario	Objetivo claro
1	3	4	3	1
2	3	4	3	4
3	4	4	4	3
Media	**3,33**	**4**	**3,33**	**2,67**

Tabla 30: Evaluación obtenida en la primera fase

Los resultados de las evaluaciones se han recogido en el siguiente diagrama multidimensional, donde la figura de color rojo representa la media de todas las evaluaciones:

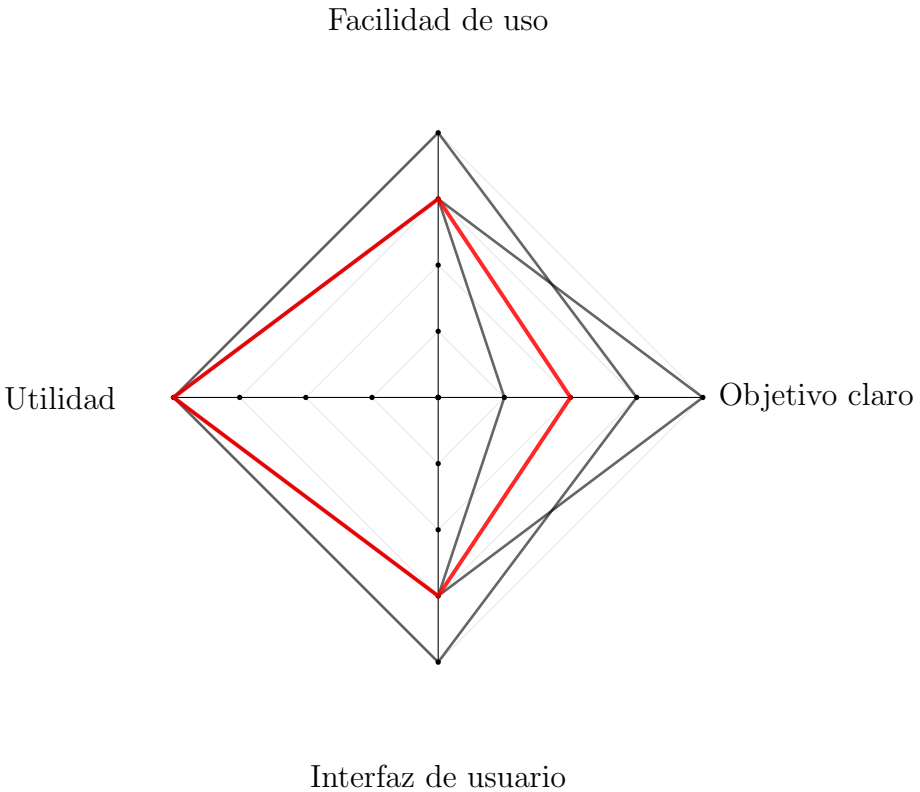


Figura 37: Diagrama multidimensional de las evaluaciones

De esta primera fase se han obtenido algunas conclusiones. En primer lugar, algunas de las personas evaluadas no sabían interactuar con los mapas que ofrecen Google Maps. Sin embargo, después de averiguar cómo funcionan, no han tenido mayores complicaciones para usarlo.

Además, las personas más jóvenes desconocían que utilidad podía tener una herramienta como esta. Esto puede deberse a la falta de información que se les había dado.

Aquellos evaluadores ajenos al ámbito de la informática preguntaron el significado de la expresión ïnterfaz de usuario", concepto que hubo que explicarles.

Teniendo en cuenta estos factores, se decidió realizar una segunda fase de evaluación con algunas diferencias.

5.2.2. Fase 2

Para realizar la segunda fase de evaluación se ha partido del método propuesto por Lam et al. [33]. Los autores identifican siete distintos escenarios para la evaluación de visualización de información:

1. Entornos y prácticas de trabajo
2. Análisis visual de datos y razonamiento
3. Comunicación a través de visualización
4. Análisis de datos colaborativo
5. Rendimiento del usuario
6. Experiencia del usuario
7. Evaluación automatizada de visualizaciones

Puesto que el objetivo de esta herramienta es transmitir información a través de la visualización de tweets, se ha centrado la evaluación en el escenario número 3: comunicación a traves de visualización. Esta evaluación estudia si y cómo se puede realizar la comunicación a través de la visualización, considerando comunicación al aprendizaje, enseñanza o presentación de una idea. Para ello, se proponen las siguientes preguntas:

- ¿Aprenden los usuarios mejor o más rápido utilizando la herramienta de visualización
- ¿Es útil la herramienta a la hora de transmitir información?
- ¿Puede extraerse información útil de la visualización de la información?

Se han realizado estas preguntas a un grupo de cuatro potenciales usuarios y se les ha pedido que respondan con una puntuación entre 0 y 4 - siendo 0 la mínima y 4 la máxima. Sus respuestas se recogen a continuación:

Evaluador	**Pregunta 1**	**Pregunta 2**	**Pregunta 3**
1	4	4	4
2	4	3	2
3	4	3	3
4	4	3	3
Media	**4**	**3,25**	**3**

Tabla 31: Preguntas de evaluación de comunicación a través de visualización

Para evaluar la experiencia de usuario, se les han preguntado también las siguientes cuestiones:

- ¿Qué es lo que te parece más útil?
- ¿Qué crees que le falta?
- ¿Qué limitaciones opinas que tiene?
- ¿Te parece fácil de entender y de aprender a usar?

Exceptuando la última pregunta, todos los entrevistados tuvieron dificultades para responder a estas preguntas, ya que no sabían que responder, por lo que solamente esta última se ha tenido en cuenta para la evaluación. Las respuestas obtenidas han sido las siguientes:

Evaluador	**Pregunta 4**
1	3
2	3
3	3
4	4
Media	**3,25**

Tabla 32: Pregunta de evaluación de la experiencia de usuario

Los resultados de estas cuatro preguntas que componen la segunda fase de evaluación del proyecto se han recogido en el siguiente diagrama multidimensional, donde la figura de color rojo de nuevo representa la media de todas las evaluaciones:

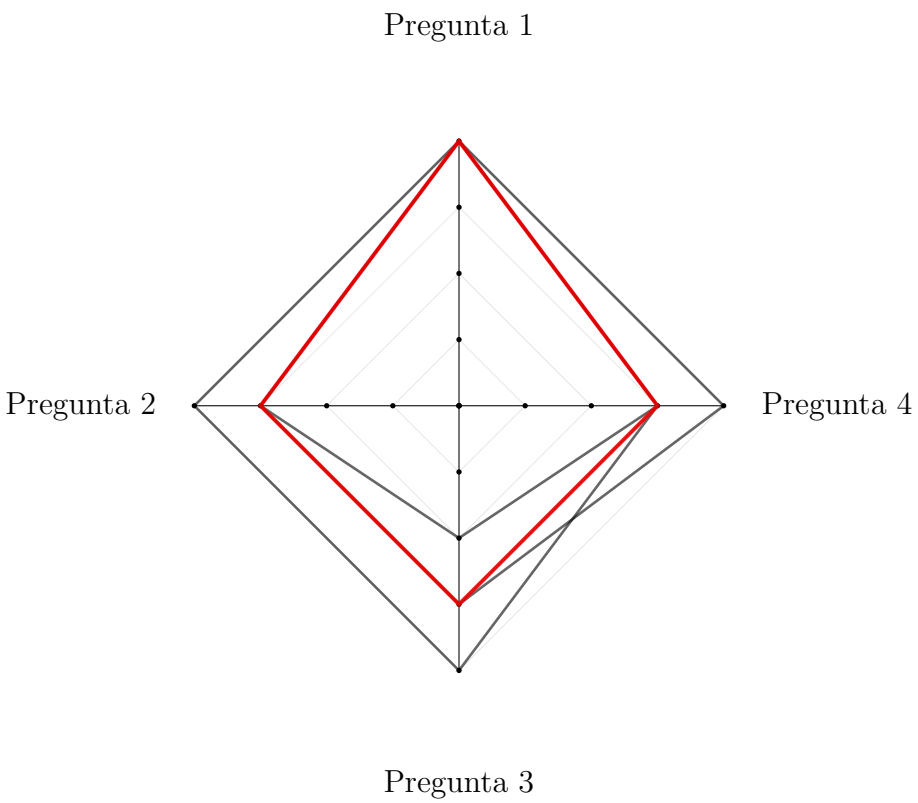


Figura 38: Diagrama multidimensional de las evaluaciones

6. Planificación del trabajo

A lo largo de esta sección, se describe cómo ha sido la planificación del trabajo.

6.1. Tareas principales

En este proyecto se identifican las siguientes tareas principales:

- **Investigación**: Leer artículos que ya han sido publicados sobre trabajos similares, con el fin de comprender la situación actual y conocer las distintas posibilidades y oportunidades que existen.
- **Extracción de datos**: Extraer los datos de Twitter con los que se va a trabajar y almacenarlos en una base de datos.
- **Análisis de términos**: Se deben analizar los datos para extraer aquellos más importantes, y almacenarlos, junto con otra información relevante.
- **Visualización**: Visualización en un mapa de los tweets.
- **Evaluación**: Evaluación del proyecto.
- **Memoria**: Redacción de la memoria del trabajo.

6.2. Sub-tareas

Habiendo identificado todas las tareas principales para este proyecto, se ha dividido cada una en sub-tareas menores, dando lugar a la siguiente lista de tareas y sub-tareas:

- **1. HERRAMIENTA VISUAL PARA MONITORIZACIÓN DE TWEETS DURANTE EVENTOS EXCEPCIONALES**
  - **2. Investigación**
    - 3. Estudiar publicaciones sobre trabajos anteriores sobre las redes sociales y los servicios de emergencias
    - 4. Estudiar distintos métodos de visualización de tweets
    - 5. Estudiar publicaciones visualización de tweets y servicios de emergencias
  - **6. Extracción de datos**
  - **7. Análisis de términos**
    - 8. Almacenar tweets en ficheros de texto
    - 9. Extraer y almacenar sustantivos en la base de datos

- 10. Categorizar términos almacenados

- **11. Visualización**
  - 12. Mostrar mapa
  - 13. Añadir marcadores de prueba
  - 14. Añadir marcadores utilizando las coordenadas de los tweets de la base de datos
  - 15. Mostrar ventanas de información con el contenido de los tweets
  - 16. Clustering
  - 17. Representar clusters con polígonos
  - 18. Cambiar color del cluster en función del término más frecuente
  - 19. Mostrar etiqueta con término más frecuente
  - 20. Añadir animación
- **21. Evaluación**
- **22. Memoria**

6.3. Estructura de descomposición del trabajo

Con el fin de facilitar la comprensión visual de la tarea, se proporciona el siguiente WBS (Work Breakdown Structure) o Estructura de descomposición del trabajo (Figura 39):

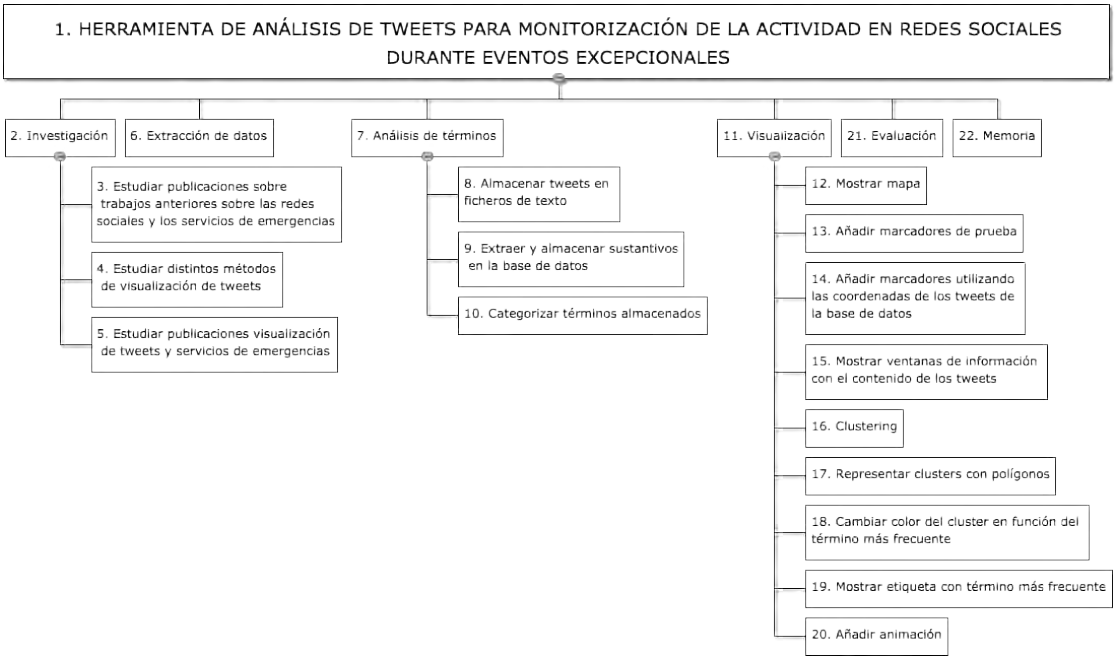


Figura 39: Estructura de descomposición del trabajo

6.4. Estimación de tiempo

Para cada una de las tareas y sub-tareas de este proyecto, se ha estimado el tiempo necesario para completarlas, expresado en días. Se ha considerado únicamente los días

Id	Descripción de la tarea	Estimación Optimista	Estimación Pesimista	Tiempo real
1	**HERRAMIENTA VISUAL PARA MONITORIZACIÓN DE TWEETS**	**120**	**198**	**156**
2	**Investigación**	**12**	**35**	**23**
3	Estudiar publicaciones sobre trabajos anteriores sobre las redes sociales y los servicios de emergencias	5	15	9
4	Estudiar distintos métodos de visualización de tweets	2	5	4
5	Estudiar publicaciones visualización de tweets y servicios de emergencias	5	15	10
6	**Extracción de datos**	**2**	**10**	**4**
7	**Análisis de términos**	**13**	**30**	**34**
8	Almacenar tweets en ficheros de texto	1	5	2
9	Extraer y almacenar sustantivos en la base de datos	2	10	11
10	Categorizar términos almacenados	10	15	21
11	**Visualización**	**70**	**99**	93
12	Mostrar mapa	2	5	1
13	Añadir marcadores de prueba	5	10	32
14	Añadir marcadores utilizando las coordenadas de los tweets de la base de datos	5	10	5
15	Mostrar ventanas de información con el contenido de los tweets	5	10	1
16	Clustering	5	15	7
17	Representar clusters con polígonos	15	20	25
18	Cambiar color del cluster en función del término más frecuente	2	9	5
19	Mostrar etiqueta con término más frecuente	3	5	6
20	Añadir animación	10	15	11
21	**Evaluación**	**12**	**12**	**12**
22	**Memoria**	**20**	**41**	**34**

Tabla 33: Estimación de días

6.5. Calendario de trabajo

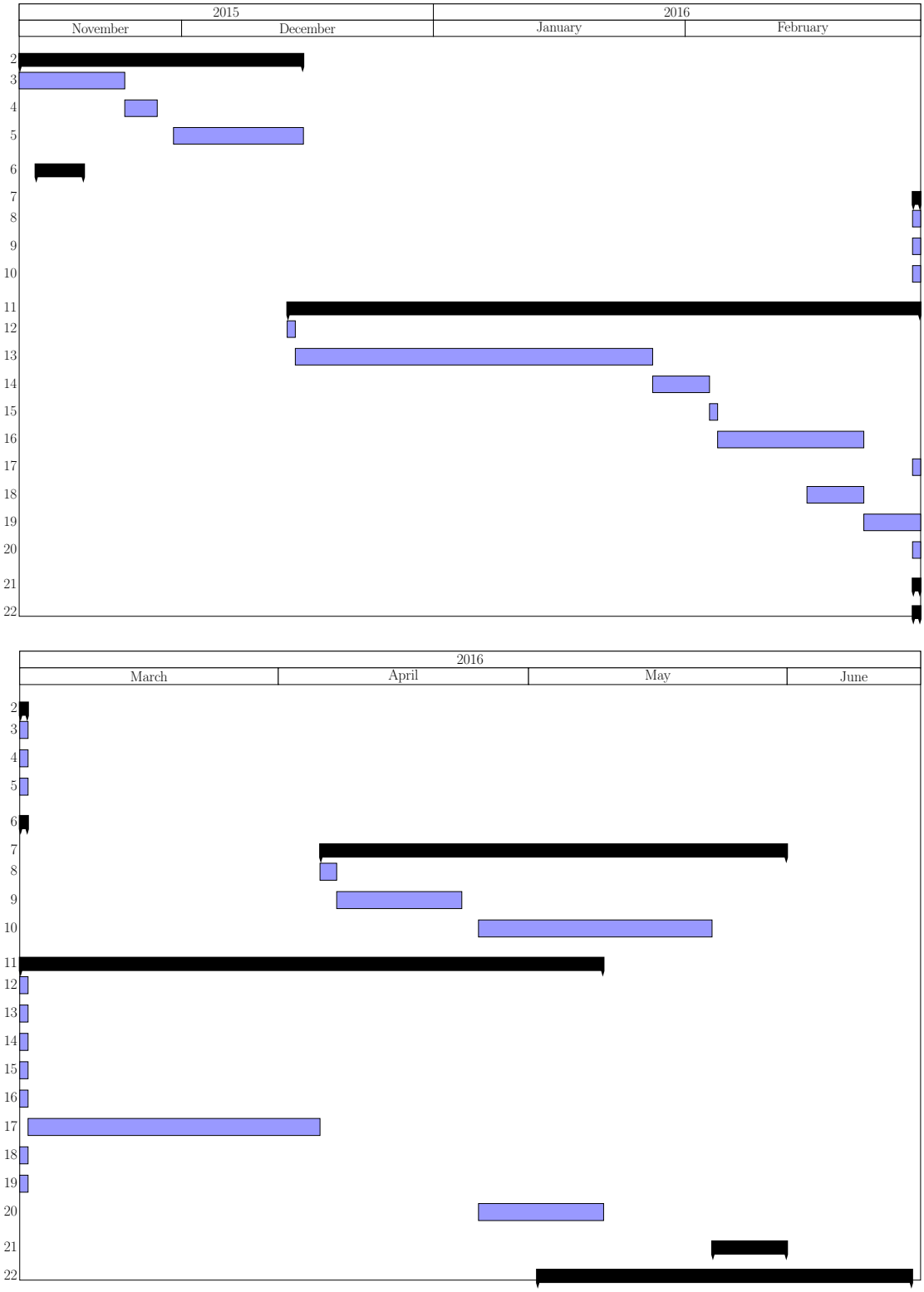
La tabla a continuación muestra la fecha real de inicio y de fin de cada tarea.

Id	Tiempo	Fecha de inicio	Fecha de fin
1	**156**	**11/11/15**	**15/6/16**
2	**23**	**11/11/15**	**11/12/15**
3	9	11/11/2015	23/11/2015
4	4	24/11/2015	27/11/2015
5	10	30/11/2015	11/12/2015
6	**4**	**13/11/2015**	**18/11/2015**
7	**34**	**06/04/2016**	**31/05/2016**
8	2	06/04/2016	07/04/2016
9	11	08/04/2016	22/04/2016
10	21	25/04/2016	22/05/2016
11	**93**	**14/12/2015**	**09/05/2016**
12	1	14/12/2015	14/12/2015
13	32	15/12/2015	27/01/2016
14	5	28/01/2016	03/02/2016
15	1	04/02/2016	04/02/2016
16	7	05/02/2016	15/02/2016
17	25	02/03/2016	05/04/2016
18	5	16/02/2016	22/02/2016
19	6	23/02/2017	01/03/2016
20	11	25/04/2016	09/05/2016
21	**7**	**23/05/2016**	**31/05/2016**
22	**34**	**02/05/2016**	**15/06/2016**

Tabla 34: Calendario de trabajo

6.5.1. Diagrama de Gantt

Con estas fechas, se ha generado el siguiente digrama de Gantt:



7. Presupuesto

En esta sección se detalla el presupuesto que supondría la realización de este proyecto.

7.1. Personal

Este trabajo puede ser realizado por una única persona. Como se ha indicado en el apartado anterior, se requiere un total de 153 días, durante los que se estima una dedicación de entre dos y tres horas diarias, lo supone una cantidad aproximada de **382 horas** trabajadas.

Si se le paga 16,5€- por hora - de acuerdo con la media del país [34] - resulta en **6.303€** de gastos de personal para la realización del proyecto.

7.2. Equipos

Para la implementación de la herramienta, se ha empleado un ordenador MacBook Air de 11 pulgadas. Actualmente, puesto que la Universidad Carlos III de Madrid forma parte del programa Apple On Campus, podrá ser adquirido por un precio de 966,79€[35].

No obstante, este equipo podrá ser empleado para la realización de distintos proyectos, por lo que el coste real deberá dividirse entre la totalidad de estos proyectos. Con esto, se consigue amortizar parte de los gastos.

La herramienta deberá estar alojada en un servidor. Para esta labor se ha elegido alojar la infraestructura en Amazon AWS, ya que permite un pago por uno, sin un compromiso de tiempo, y al ser un servicio cloud nos da mayor flexibilidad para aumentar o disminuir los recursos necesitados en caso necesario.

El coste de un servidor Linux t2.medium de doble núcleo alojado en Irlanda es de 0.056$ por hora [36]. Por contrato, el precio mensual es de 29,20$ al mes, lo que supone un total de 350,4$ anuales.

Se deberá contratar también servicio de almacenamiento para la base de datos. El precio de 5 instancias de base de datos MySQL con un crecimiento de 5GB mensual por instancia es de 4315,32$ mensuales. Esto hace un total de 51783,84$ anuales.

Sería conveniente contratar distintas isntancias de bases de datos, para así poder dedicar cada una de éstas a los tweets de un evento diferente, y poder analizar más de uno.

7.3. Resumen

Es necesario un pago inicial para adquirir el ordenador necesario para la implementación de la herramienta.

Durante los siete siguientes meses, es necesario pagar al ingeniero responsable del desarrollo de la herramienta, al que se le pagará **900,43€** al mes. Se pagará tambien 29,20$ por el servidor y 4315,32$ por la base de datos.

A partir de estos siete meses, ya únicamente será necesario pagar 29,20$ por el servidor y 4315,32$ por la base de datos.

En la tabla a continuación se recogen los gastos durante los doce primeros meses del proyecto:

Mes	Gastos
1	5706,99 €
2	4740,2 €
3	4740,2 €
4	4740,2 €
5	4740,2 €
6	4740,2 €
7	4740,2 €
8	3839,77 €
9	3839,77 €
10	3839,77 €
11	3839,77 €
12	3839,77 €

Tabla 35: Gastos durante los 12 primeros meses del proyecto

Se ha estimado el precio a euros basándose en su correspondencia a día de 9 de junio de 2016, momento en el que 1€equivale a 1,13$.

Estos datos se reflejan también en la siguiente gráfica:

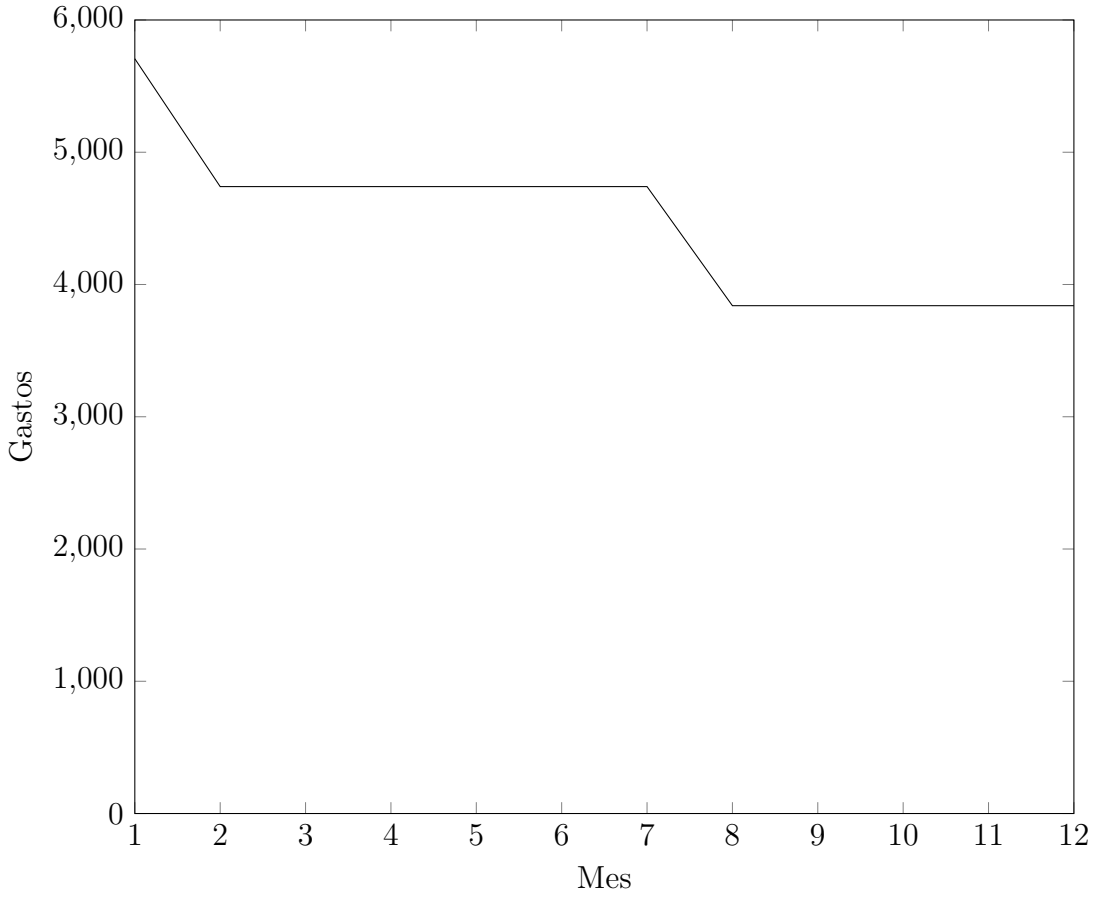


Figura 40: Gastos durante los 12 primeros meses del proyecto

8. Conclusiones

La popularidad de las redes sociales está en auge. Con tal cantidad de información generándose cada segundo, es necesario encontrar una forma de analizar y dar sentido a todas estas publicaciones. El objetivo de este Trabajo de Fin de Grado ha sido desarrollar una herramienta para el análisis visual de las tweets - publicaciones de la red social Twitter - para la monitorización de la actividad de redes sociales durante eventos excepcionales que pueda ser de utilidad para los servicios de emergencias.

Con esto, no se pretende reemplazar ni sustituir los métodos que ya existen, como las llamadas telefónicas para pedir ayuda, sino dar un paso más y aportar algo más que permita una mejor comprensión de cómo los ciudadanos reaccionan y piden ayuda durante ciertos sucesos.

Uno de los problemas con los que se ha tenido que lidiar durante el desarrollo de la aplicación ha sido la falta de datos con geolocalización en Europa. Mientras que en los trabajos previos de otros investigadores - basados en eventos ocurridos en Norteamérica - no se menciona ningún problema con respecto a esto, de los datos recogidos de los atentados de París, muy pocos tweets tenían coordenadas. Además, la mayoría de los tweets que sí tenían esta opción activada, se situaban en Norteamérica o Asia, de lo que se deduce que la población europea es más reacia a compartir su ubicación.

No obstante, considero que la herramienta propuesta cumple su objetivo y puede ayudar a los operarios de servicios de emergencias, por ejemplo, a identificar zonas afectadas. Además, como se señaló en el Estado del Arte, no se han podido encontrar otras herramientas con la funcionalidad y objetivos de la que se ha desarrollado para este Trabajo de Fin de Grado, por lo que cubre esta necesidad que hasta ahora no estaba cubierta.

En futuros trabajos, sería interesante desarrollar la herramienta en tiempo real. La herramienta ya se ha preparado para esto mediante la animación que se le ha añadido. Ésta muestra cómo sería la visualización si los tweets se agregasen progresivamente, como sería en tiempo real. Además, permite procesar cantidades incrementales de tweets.

Preparar este tipo de herramienta para que funcione en tiempo real supone esperar a que ocurra algún evento para poder probarla. Estos eventos nunca se sabe cuándo van a ocurrir, por lo que es difícil estar preparados para ellos. Por esto, no se ha podido desarrollar en tiempo real para este proyecto.

A nivel personal, desarrollar este proyecto ha sido una experiencia realmente interesante. En primer lugar, me ha permitido trabajar con herramientas que no había tratado en la carrera, como por ejemplo PHP. También ha sido una gran toma de contacto con el ámbito de la investigación, en el que siento que no se trabaja mucho durante los años de carrera.

Antes de empezar este Trabajo de Fin de Grado, desconocía cómo era la labor de los investigadores. Sin embargo, durante este año he leído artículos científicos, comparado alternativas de diseño e incluso trabajado en la redacción de un artículo junto con la tutora de este trabajo.

Pero sin duda, lo que más me ha aportado este trabajo ha sido la realización de cuan útil puede ser el ámbito de la informática y las ciencias de la información en situaciones tan delicadas como estos sucesos. En un futuro, me gustaría poder continuar trabajando en proyectos similares, y que los conocimientos adquiridos durante estos años de universidad puedan servir para ayudar o simplificar la labor de otras personas.

Referencias

[1] F. Bracero, "La edad media de acceso de los menores al móvil es a los 13 años," 2012. [Online]. Disponible: http://www.lavanguardia.com/tecnologia/20120618/54313198104/edad-media-acceso-menores-movil-13-anos.html

[2] S. Kemp, "Digital in 2016," 2016. [Online]. Disponible: http://goo.gl/7PU9ln

[3] Facebook, "Company info," 2016. [Online]. Disponible: http://newsroom.fb.com/company-info/

[4] Worldometers, "China Population," 2016. [Online]. Disponible: https://aws.amazon.com/es/ec2/pricing/

[5] ——, "World population," 2016. [Online]. Disponible: http://www.worldometers.info/world-population/china-population

[6] Internetlivestats, "One second," 2016. [Online]. Disponible: http://www.internetlivestats.com/one-second/#tweets-band

[7] Twitter, "Empresa About," 2016. [Online]. Disponible: https://about.twitter.com/es/company

[8] T. Finin and B. Tseng, "Why we twitter: understanding microblogging usage and communities," ACM, 2007, pp. 56–65.

[9] A. L. HughesL. Palen, "Twitter adoption and use in mass convergence and emergency events," *International Journal of Emergency Management*, vol. 6, no. 3-4, pp. 248–260, 2009.

[10] P. Treleaven and H. E. Stanley, "Quantifying the digital traces of hurricane sandy on flickr," *Sci. Rep*, vol. 3, p. 3141, 2013.

[11] T. OnoratiP. Díaz, "Semantic visualization of twitter usage in emergency and crisis situations," Springer, 2015, pp. 3–14.

[12] P. y. K. L. y. K. G. y. D. M. y. K. B. Bhulai, Sandjai y Kampstra, "Trend visualization on twitter: What's hot y what's not?" *Data analytics*, pp. 43–48, 2012.

[13] H. Liu and R. Maciejewski, "Understanding twitter data with tweetxplorer," ACM, 2013, pp. 1482–1485.

[14] Y. Hu and S. C. North, "Interactive visualization of streaming text data with dynamic maps." *J. Graph Algorithms Appl.*, vol. 17, no. 4, pp. 515–540, 2013.

[15] D. H. y Markus Schedl y Andrej Košir y Marko Tkalčič, "The million musical tweets dataset: What can we learn from microblogs," , , November 2013.

[16] C. ChewG. Eysenbach, "Pandemics in the age of twitter: content analysis of tweets during the 2009 h1n1 outbreak," *PloS one*, vol. 5, no. 11, p. e14118, 2010.

[17] J. Li, Yue y Shan, "Understanding the spatio-temporal pattern of tweets," *Photogramm. Eng. Remote Sens*, vol. 79, pp. 769–773, 2013.

[18] K. y. S. Y. y. K. B. Meyer, Brett y Bryan, "Twitterreporter: Breaking news detection y visualization through the geo-tagged twitter network." , 2011, pp. 84–89.

[19] G. y. A. M. A. y. L. H. Kumar, Shamanth y Barbier, "Tweettracker: An analysis tool for humanitarian y disaster relief," , 2011.

[20] M. Naaman and F. Kivran-Swaine, "Diamonds in the rough: Social media visual analytics for journalistic inquiry," IEEE, 2010, pp. 115–122.

[21] L.-E. Haug and M.-C. Hsu, "Visual sentiment analysis on twitter data streams," IEEE, 2011, pp. 277–278.

[22] X. Zhang and J. Blanford, "Senseplace2: Geotwitter analytics support for situational awareness," IEEE, 2011, pp. 181–190.

[23] S. Liu and H. Qu, "Whisper: Tracing the spatiotemporal process of information diffusion in real time," *IEEE Transactions on Visualization and Computer Graphics*, vol. 18, no. 12, pp. 2649–2658, 2012.

[24] B. Robinson and R. Power, "Using social media to enhance emergency situation awareness," *IEEE Intelligent Systems*, vol. 27, no. 6, pp. 52–59, 2012.

[25] I. Aedo and F. Astorga-Paliza, "Sema4a: An ontology for emergency notification systems accessibility," *Expert Systems with Applications*, vol. 37, no. 4, pp. 3380–3391, 2010.

[26] M. Liu and F. Wu, "Opinionflow: Visual analysis of opinion diffusion on social media," *IEEE transactions on visualization and computer graphics*, vol. 20, no. 12, pp. 1763–1772, 2014.

[27] Twitter, "Developer Agreement & Policy," 2016. [Online]. Disponible: https://dev.twitter.com/overview/terms/agreement-and-policy

[28] O. de las Naciones Unidas, "Declaración Universal de los Derechos Humanos," 2016. [Online]. Disponible: http://www.un.org/es/documents/udhr/

[29] BOE, "Documento consolidado BOE-A-1999-23750," 2016. [Online]. Disponible: https://www.boe.es/buscar/act.php?id=BOE-A-1999-23750

[30] J. Perez Serna, "La API de Twitter y la Ley Orgánica de Protección de Datos (LOPD)," 2016. [Online]. Disponible: http://estwitter.com/2010/07/02/la-api-de-twitter-y-la-ley-organica-de-proteccion-de-datos-lopd/

[31] T. S. N. L. P. Group, "The Stanford Natural Language Processing Group," 2016. [Online]. Disponible: http://nlp.stanford.edu/software/tagger.shtml

[32] P. University, "About Wordnet," 2016. [Online]. Disponible: https://wordnet.princeton.edu/

[33] E. y. I. P. y. P. C. y. C. S. Lam, Heidi y Bertini, "Seven guiding scenarios for information visualization evaluation," 2011.

[34] Eurostat, "Hourly labour costs," 2016. [Online]. Disponible: http://ec.europa.eu/eurostat/statistics-explained/index.php/Hourly{_}labour{_}costs

[35] Apple, "Compra un MacBook Air," 2016. [Online]. Disponible: http://www.apple.com/es{_}aoc{_}5005152/shop/buy-mac/macbook-air

[36] Amazon, "Precios de Amazon EC2," 2016. [Online]. Disponible: https://aws.amazon.com/es/ec2/pricing/

Anexo: English Summary

Introduction

The aim of this Thesis is to implement a visual tool for monitoring social network activity during exceptional events. In this section of the document, the motivation behind this project is explained, as well as the goals we expect to achieve, the initial existing problem and the solution suggested to overcome this problem.

Socio-economic environment

In the last few years, there has been a rise in the use of new technologies. We live in a society where a person acquires their first phone at an average age of 13 years old [1], and 80 % of our country's population owns a smartphone [2].

As opposed to home phones, mobile phones allow us to stay in contact with other people wherever and whenever we need. This has a huge impact on our personal lives and in our work environments.

Besides, smartphones come with a wide range of applications: from games to keep your entertained, to a calculator, notepad, o music players. Another great advantage these devices offer is a connection to the internet, which makes keeping in touch even simpler. For example, via social networks.

We have witnessed a rise in the use of these too. These platforms provide a communication channel for a large number of users. In social networks, one can find very different kind of messages. They can be used to share what we are doing, or to share our opinions and feelings. It is also common to share news on them.

The appearance of social networks has made communication easier in many ways. On the one hand, they offer a communication channel that is easy to use and available at any moment. On the other hand, most of these networks make it easy for their users to find other users that they may know, such as old classmates, colleagues, relatives, o even accounts associated with phone numbers found in the user's phone.

These networks also allow us to share new content and approach people that we do not know personally. For instance, posts can be shared publicly, in a way that any people that find that post - either they are a user of the social network where it was shared or not - can read them. Therefore, this allows users to reach a large audience.

The impact of these platforms in our society is certainly interesting. For example, the social network Facebook counts with 1,65 billions of monthly active users [3], out of which, 1,51 billion are active through their cellphones. This figure

is even greater than the population of the most populated country on the planet, China, which has 1,38 billions of inhabitants [4]. Thus, were Facebook a country, it would be the largest one in the world. Considering that the world's population is 7.4 billions of people [5], 2 out of 7 of these would be from the country Facebook. These numbers prove the high penetration rate of social networks in today's society.

On the other hand, during the current year (2015-2016), Europe been the target of two terrible jihadist attacks in the capital cities of France and Belgium. This type of attacks is nothing new. In fact, they have been occurring frequently for the past years. Back in 2001, the city of New York was the target of one of them, not to mention the attacks of 2004 in Madrid. A threat that worries a big part of the population and that, as such, generates a huge response in social networks, where the activity increases every time an event of this kind happens.

It is common for the citizens to come to these networks with different aims: seeking information, to try to contact their loved ones, to express their concern, or to show their support for the victims.

Nevertheless, these events are not the only ones that have an impact like this on social networks. Other events with similar effects are natural catastrophes - like the hurricane Sandy, that hit the United States of America in 2014 - or special celebrations, like some sports events.

Motivation

As it has been previously mentioned in this document, there is a huge amount of users that come to social networks to share all kinds of content. What's more, activity on these increases during exceptional circumstances.

It would be interesting for an emergency system operator to be able to comprehend what is going on in the population when one of these events takes place. When there is an accident or an emergency, it is vital for this operator to be able to respond as fast as possible. In order to be able to do this, they should have all the necessary and updated information. With so many users sharing so much content, social networks become a place from which we can retrieve first-hand information, directly from the population.

The problem

Even though this information can indeed be found in social networks already, it is necessary to process it and visualize it in an efficient way. In a single second, over 7.000 tweets [6] are shared. It would be unimaginable for an operator to read all these posts that are being shared in a single instant, since what they are seeking for is a fast response. Having to stop to read all this information would take too long and would keep the operator from responding on time.

This introduces us to our next challenge: how can this amount of information be used to help the emergency services?

Goals

The goal of this Thesis is to offer a tool that will allow emergency services to use all this large amount of information that is published on Twitter during an exceptional event, and analyze it in a way that it will become meaningful and useful to the operators of these services.

With this goal in mind, a visualization tool has been developed that allows the monitorization tweets during exceptional events.

Such tool aims to achieve the following specific goals:

- Retrieve messages shared on social networks about a specific topic.
- Analyze the retrieved data semantically, categorizing the terms by their relevance regarding the chosen topic.
- Visuzalize the data in an efficient way, so that the users - and in particular, the operators of the emergency services - can make use of them.
- Use the proposed tool with a real-life use case.

This way, we can go a step further in the traditional performance of an emergency system, using already existing information that is obtained directly from the population, to improve the comprehension of the scope and effect of these events.

Final solution

To solve the existing problem - finding a way to visualize the large amount of data that can be retrieved from user's publications on social networks - the following tweet visualization tool is proposed.

Twitter is a micro-blogging platform that was created in 2006, whose users can read and share messages with a maximum of 140 characters. Access to the social network is allowed either through their website or through their apps for smartphones and tablets. This means their users can contribute with new content easily, using different devices, and at any moment.

To understand the scope of this platform, it should be mentioned that Twitter counts with 310 million monthly active users [7]. In a single second, an average of

7.203 tweets is published [6]. These numbers prove how a large part of our society is active in this network, where new content is being written frequently.

Twitter users can also choose between creating a public profile - that way anyone, registered or not, can read their posts - or private - so that only authorized users can read them. This means that a lot of tweets are public. For these reasons, it has been decided that the data used for this tool is retrieved from Twitter.

Implementation of the tool

The proposed tool is designed to monitor the activity generated in social media during exceptional events. In the following sections, the implementation of this tool will be described in detail.

How it works

This tool uses a compilation of tweets stored in a database, collected during an exceptional event. The content of these tweets is analyzed syntactically using Stanford POS (Part Of Speech) Tagger, in order extract the nouns and store them in our database. These are later analyzed semantically, categorizing them according to their relevance.

The tool then accesses the database using SQL queries and orders the stored tweets by their publication date in an ascending order. The needed information about the tweets and the nouns is stored in array structures that the javascript script will use for the visualization part.

For the visualization of the map, Google Maps API has been used. The tool first loads a map of the area affected by the events and progressively adds a marker for each tweet with coordinates found in the database. Every time a marker is added, the tool calculates the possible clusters - i.e. groups of tweets that are close geographically - and a polygon that would cover all the markers in the each cluster. For this, the Graham Scan - which will be described later in this document - is used. For each cluster, a label with the most common term found in the tweets that belong to that cluster is shown. This term determines the color of the polygon representing the cluster, as each of the different categories, a term can belong to has a different color associated with it.

Data

Initially, the data has been retrieved from Twitter using Twitter's Search API and stored into a database that followed the structure specified by Twitter. Nevertheless, some of these tables are not necessary for our tool, and can, therefore, be removed from our database. We also need to create additional tables to store all the information about the terms found in the tweets. As a result of these two issues, the structure of the final database is shown in the following Enhanced Entity-Relationship Diagram (EER) (Figure 41):

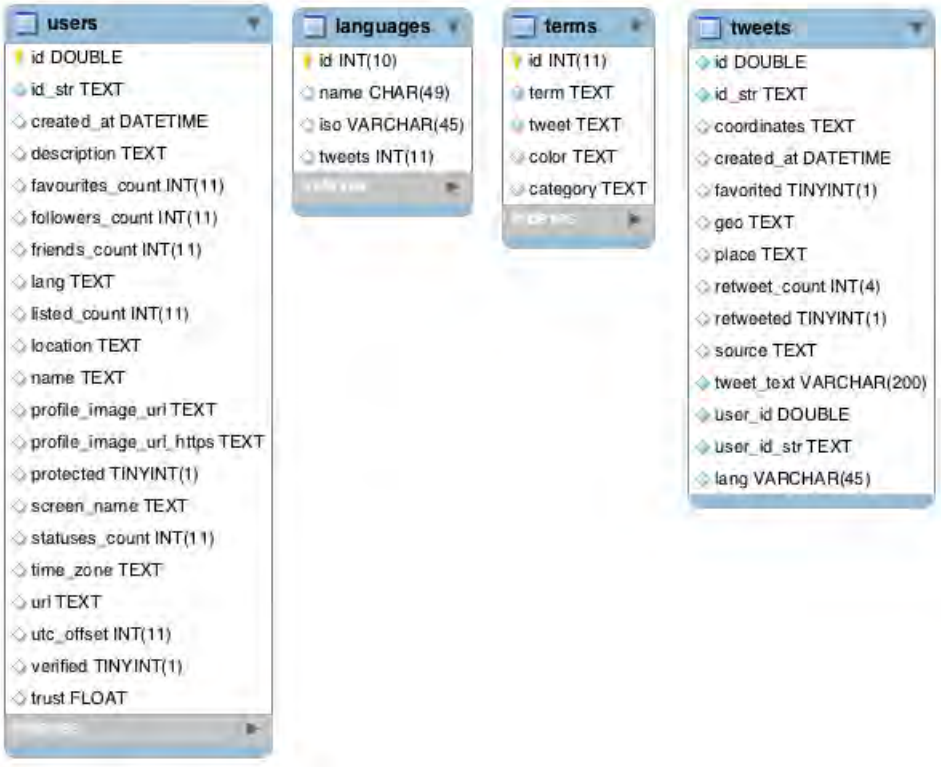


Figura 41: EER of the database

The tweets are then semantically analyzed using Stanford POS (Part-of-Speech) Tagger [31], that assigns parts of speech - a category to which a word is assigned in accordance with its syntactic functions - to each word. This way we are able to extract the nouns from the tweets content and store them in the database.

The terms are later analyzed semantically and categorized into one of the following categories:

- **Emergency**: terms from an emergency domain
- **Time**: time expressions
- **Place**: list of countries.
- **Media**: terms regarding the media
- **Evacuation**: terms from an evacuation domain

To do this, an ontology [25] and two taxonomies are used. With the following ontology (Figure 42) we categorize the terms belonging to the Emergency, Media, and Evacuation categories.

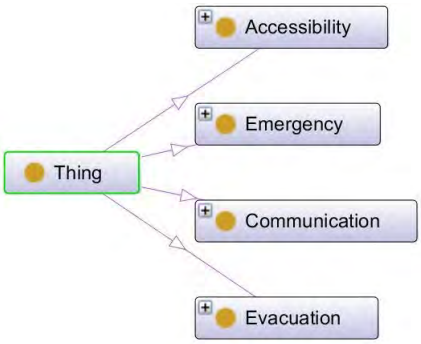


Figura 42: Ontology

WordNet is a lexical database that groups English words into sets of synonyms. Using WordNet [32], the tool checks if the term is a synonym of either the word *place* or *time*. If that is the case, the term is assigned the category it is a synonym of. Otherwise, two different taxonomies are used in order to determine if the term belongs to either the Time or Place categories: one with time expressions, and one with a list of countries, respectively. The terms that do not fit into any of the previously mentioned categories are assigned the category General.

Visualization

For the visualization of the maps, Google Maps API has is used. This API allows us to easily load a map and add different objects to it. One of these objects is a **marker**. Markers are used to represent locations on a map. In this tool, a marker is used to represent a tweet stored in the database. The content of each tweet can be seen by clicking on the marker using a **infowindow** (Figure 43).

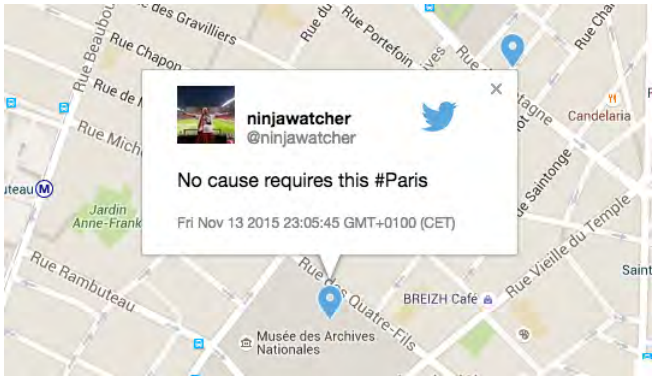


Figura 43: Infowindow

Every time a marker is added, the script calculates the sets of clusters - groups of markers with nearby coordinates - using the **MarkerClusterer** library. The library, however, is only used to calculate the different clusters and tweets in each cluster and not for the visualization part, since it does not define what the boundaries of a cluster are exactly.

Insted, Graham Scan is used instead. This algorithm calculates, given a set of points, the convex polygon with the minimum area that covers all of the points in the given set (Figure 44).

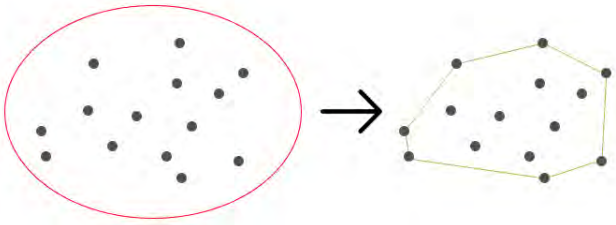


Figura 44: Convex Hull Example

In order implement this algorithm, a new function is declared, which, given an array of coordinates, returns an array structure with the coordinates that represent the vertex of the convex hull polygon.

This set of points is later represented on the map using a Google Maps **polygon** object. The color of the polygon is determined by the category of the most frequent term used in the tweets that belong to this cluster. This term is also shown using a **label** (Figure **??**).

The clusters - along with their polygons and labels - are removed every time the user zooms in or out or navitagtes to a different part of the map and recalculated, in order to ensure an optimum visualization of the data at every moment.

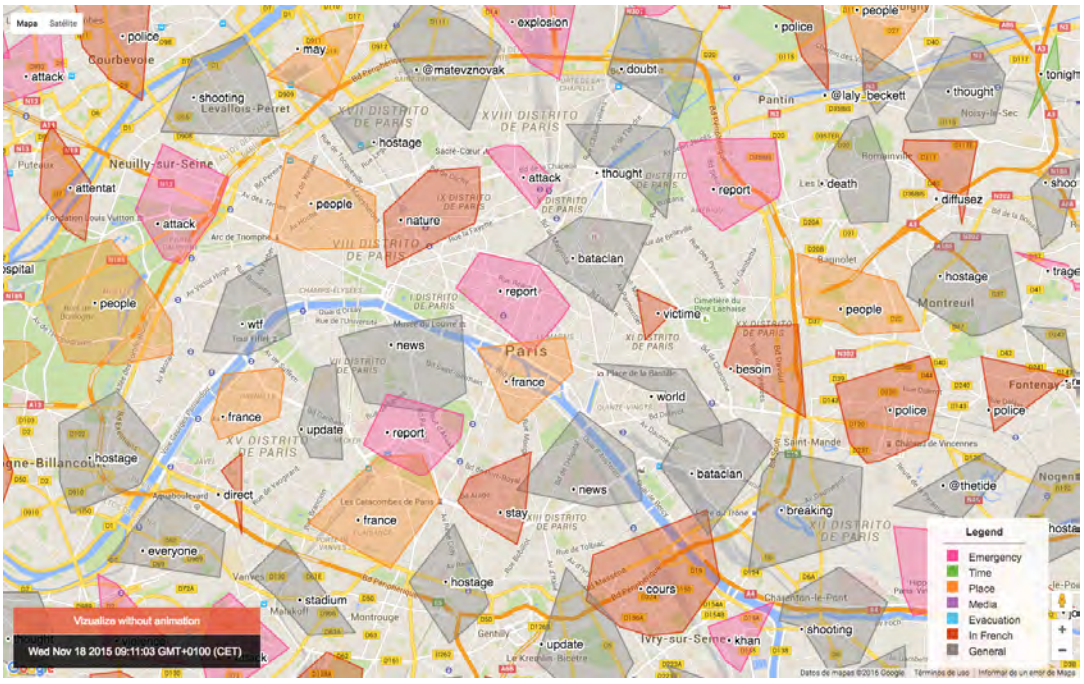


Figura 45: Final aspect of the tool

Use case

The use case used during this project has been the attacks that took place in Paris on the evening of 13 November 2015. This event was chosen due to its closeness, geographically speaking - since France is a neighboring country of Spain - as well as in time - as it happened during this year.

It is also an event that had a huge impact on Spanish society. Maybe precisely because of this geographical proximity, or maybe because Spain was also the target of a similar event 2004 Madrid train bombings.

The information retrieval process actually consisted on three different sub-processes, each of them corresponding a different query:

- **Paris**: The first query simply collected those tweets containing the word *Paris*.
- **Fusillades**: Being an event that took place in France, it was decided to obtain tweets written by the French population. To do this, terms containing the term *fusillades* - *shootings* - were retrieved.
- **#PorteOuverte**: This hashtag, which means *Open Doors*, was used by French citizens, who were offering a place in their homes to those seeking shelter.

Nevertheless, the tool can be used for any event that is wished to be analyzed, only needing to alter the content stored in the database.

For this particular use case, a total of 1,739 different terms were found. The most frequent ones of these were *France* (178 times), *attack* (158 times) and *people* (147 times). These numbers were collected after removing from the database the term *Paris*, as well as the hashtags used for the event. These were discarded since they were the most frequent terms in most of the clusters. Hiding them makes it easier to appreciate the difference of terms being used in different areas.

Conclusions

The popularity of social networks is booming. With such amount of information being generated every second, it is necessary to find a way to analyze and make sense of all these publications. The goal of this Thesis has been to develop a tool that allows the visual analysis of tweets - posts from the social network Twitter - for monitoring social networks activity during exceptional events, in a way that can be useful for emergency services.

By doing this, we do not intend to replace any of the existing methods, such as telephone calls asking for help, but to go one step further and contribute with a tool that will allow a better comprehension of how citizens react to certain circumstances.

One of the problems faced during the development of this application has been the lack of data with geolocation in Europe. Whereas in previous works carried out by other investigators - based on events that took place in North America - this problem is never mentioned, of all the data retrieved regarding Paris attacks, very few had coordinates. Besides, most of the tweets that did have a location were located in North America or Asia. This leads to the conclusion that the European population is less likely to share their location.

Nevertheless, I personally consider that the suggested tool fulfills its purpose and can indeed assist emergency services operators in their work. For example, by identifying the zones affected by a tragedy. Besides, after studying the State of the Art, no other tools with these function and goals have been found. Therefore, the proposed tool that was developed for this Thesis fills this existing need.

For future works, it would be interesting to implement the tool in real-time. The tool has already been prepared for this, by adding the animation effect. This shows what the visualization would be like if tweets were being added progressively, as would be the case if the application was working in real time. Besides, it allows processing incremental amounts of tweets.

Preparing a tool like this to work real-time means having to wait for an event to take place in order to be able to test it. Since you never know where these events will happen, it makes it extremely difficult to be ready for one of them. For this reason, it has not been possible to develop the tool in real-time for this project.

Personally, I consider developing this project a really interesting experience. In the first place, it has allowed me to work with tools I had not worked with before, such as PHP. It has also been a great introduction to the world of research, which I feel we do not work on very much during the years of University.

Before I started working on this Thesis, I did not know what the work of researchers was. However, during this year, I read many research papers, studied design alternatives, and even worked on writing a paper myself along this project's tutor.

But without a doubt, of all the things this project has given me, what I value the most has been the realization of how useful the field of computer science and information technology can be during situations as delicate as these events. In the future, I would love to be able to continue working on similar projects, so that all the knowledge that has been acquired during all these years at university can be used to help or simplify the tasks of other people.